Humanités numériques et science ouverte
Les collectifs

Christine Kosmopoulos & Joachim Schöpfel (dir.)

Publier, partager, réutiliser les données de la recherche : les *data papers* et leurs enjeux

La gestion, le partage et l'ouverture des données de recherche font partie des priorités de la politique de la science ouverte. Ainsi, en France, le ministère de l'Enseignement supérieur et de la Recherche a défini la structuration et l'ouverture des données de recherche comme un des axes majeurs de ses deux Plans nationaux pour la science ouverte (2018 et 2021). L'objectif du second Plan (PNSO2 : 2021-2024) est une généralisation de la science ouverte d'ici 2024 autour de plusieurs mesures, dont l'obligation de la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics et la promotion de l'adoption d'une politique d'ouverture pour les données associées aux articles publiés par les chercheurs

Septentrion
PRESSES UNIVERSITAIRES

OUTILS ET INFRASTRUCTURES

Sauf mention contraire, tous les liens ont été vérifiés au 31/01/2025.

Introduction

Christine Kosmopoulos et Joachim Schöpfel

1. La gestion, le partage et l'ouverture des données de recherche font partie des priorités de la politique de la science ouverte. Ainsi, en France, le ministère de l'Enseignement supérieur et de la Recherche a défini la structuration et l'ouverture des données de recherche comme un des axes majeurs de ses deux Plans nationaux pour la science ouverte (2018 et 2021)[1]. L'objectif du second Plan (PNSO2 : 2021-2024) est une généralisation de la science ouverte d'ici 2024 autour de plusieurs mesures, dont l'obligation de la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics et la promotion de l'adoption d'une politique d'ouverture pour les données associées aux articles publiés par les chercheurs.

2. Les *data papers*, souvent traduits en français par « articles de données » participent de ce nouvel écosystème en émergence. Publiés dans des revues à comité de lecture, ils s'appuient sur des infrastructures de services et s'accompagnent de nouveaux protocoles de standardisation dont la finalité est de faciliter la gestion et le partage des données de recherche.

Le contexte de la science ouverte

3. Partager les derniers résultats de ses recherches avec ses pairs à des fins de réutilisation est une pratique qui remonte au temps des lettres entre savants de la République des Sciences (Rentier, 2018). L'apparition des premières revues scientifiques au XVII[e] siècle, comme le *Journal des Savants* en 1665, et l'institutionnalisation des publications scientifiques marquent une étape importante dans la manière de partager et de valider les résultats scientifiques qui s'insère dans une vision cumulative et positiviste de la science. Au XX[e] siècle, l'arrivée de maisons d'édition privées spécialisées dans la publication scientifique renforce ce modèle tout en transformant les revues en produits commerciaux, un système bouleversé récemment par l'émergence des nouvelles technologies de l'information et de la communication (NTIC) et plus particulièrement de la science ouverte. Si la diffusion des savoirs, depuis la lettre manuscrite aux entrepôts de données ouvertes, a connu de nombreuses transformations, les plus saillantes sont sans aucun doute associées à la naissance de l'imprimerie et d'Internet. La création en 1991 du e-print serveur *Arxiv* par Paul H. Ginsparg incarne la rupture majeure provoquée par les NTIC d'avec les pratiques éditoriales instaurées depuis plusieurs siècles, en offrant l'accès sans restriction aux données de la recherche en physique

1. Plan national pour la science ouverte https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525.

à travers un serveur d'auto-archivage de prépublications par les chercheurs[2]. Cette étape cruciale dans la compréhension des évolutions ultérieures sur le partage des données de la recherche s'inscrivait elle-même dans des pratiques courantes de la communauté des physiciens qui consistaient à partager en amont du long processus d'évaluation et de publication des revues, leurs prépublications afin d'accélérer les échanges entre pairs tout en garantissant la paternité des découvertes.

4. Dans les années 1990, le Web apparaît comme une extraordinaire opportunité de réappropriation de la diffusion scientifique par la communauté dans une logique non commerciale. Ainsi, « les promoteurs de cette nouvelle publicisation de la science envisagent Internet comme une solution à la crise de l'édition traditionnelle et une chance pour les scientifiques de réinvestir le secteur de la publication – en partie dominée par des logiques de profit – en se passant de la médiation jusque-là incontournable des éditeurs » (Pignard-Cheynel, 2004). Les réflexions sur l'accès ouvert en recherche connaissent leur apogée avec la conférence de Budapest de 2001 (BOAI)[3] à laquelle succéderont la déclaration de Bethesda[4] sur la publication en libre accès et la déclaration de Berlin[5] sur le libre accès à la connaissance en sciences exactes, sciences de la vie et en SHS en 2003, qui officialisent le mouvement d'appropriation né de la communauté scientifique. La prise de conscience de l'impact du coût pour les bibliothèques et les institutions des abonnements aux revues et aux bouquets de revues des éditeurs commerciaux ainsi qu'aux outils bibliométriques tels que le *Web of Science* (WoS) pour évaluer les chercheurs accentue de toutes parts ce mouvement d'émancipation à l'égard des éditeurs privés. Des revues scientifiques en libre accès indépendantes se créent en même temps que des plates-formes éditoriales ainsi que des entrepôts de données et d'archives ouvertes sur le modèle d'*ArXiv*.

5. En 1996, un rapport de l'OCDE conclut que « la configuration des systèmes nationaux d'innovation, à savoir les mouvements et les relations entre l'industrie, l'État et l'université dans le développement scientifique et technologique, est un facteur économique déterminant » (OCDE, 1996). La priorité est donc donnée de favoriser l'économie du savoir considérée comme un catalyseur de la performance économique. Les agences de financement, bailleurs de fonds nationaux et européens exigent que les données de recherche financées sur fonds publics soient librement accessibles tout en nuançant selon la formule « ouvertes autant que possible, fermées autant que nécessaire ». Classiquement les revues jouent un rôle central dans la circulation des travaux et des données scientifiques et il n'est donc pas surprenant qu'à côté des publications scientifiques classiques, des articles spécifiquement destinés à valoriser la construction des données soient en train de se mettre en place.

2. ArXiv https://arxiv.org/
3. Budapest Open Access Initiative : https://www.budapestopenaccessinitiative.org/.
4. http://legacy.earlham.edu/~peters/fos/bethesda.htm
5. https://openaccess.mpg.de/Berlin-Declaration

Les principes FAIR

6. Pour faciliter l'interconnexion des outils et systèmes dans le partage de ces données, un groupe de travail comprenant conjointement des représentants du monde universitaire, de l'industrie, des organismes de financement et des éditeurs scientifiques publie en 2016 dans *Nature* une série de principes dits FAIR visant à faciliter le partage et la réutilisation des données scientifiques (Wilkinson *et al.*, 2016). En d'autres termes, ils proposent de standardiser la construction et l'organisation des données de telle sorte que celles-ci soient trouvables (*Findable*), accessibles (*Accessible*), interopérables (*Interoperable*) et réutilisables (*Reusable*).

7. En quelques années, les principes FAIR se sont imposés comme des critères d'exemplarité dans le traitement des données en matière de modèle économique, mais aussi d'éthique et de pérennité pour le développement des infrastructures liées à la science ouverte. Leur but est bien d'accélérer la recherche par l'exploitation de la puissance de l'analyse machine interconnectée tout en garantissant la transparence, la reproductibilité et l'utilité sociale. Mais, comme on le trouvera décrit dans le présent ouvrage, pour que les machines ou les humains puissent trouver et réutiliser les données, il faut aussi que les métadonnées fournissent une description riche et détaillée de ces données.

8. La documentation rédigée dans le *data paper* sur la construction des données peut être utilisée pour garantir ces exigences et contribuer au cercle vertueux d'une science ouverte reproductible, intègre et transparente tant pour les producteurs/auteurs que pour les réutilisateurs des données. Cette ambition pourrait être aussi une réponse à la crise de confiance qu'on a pu observer en biologie ou encore récemment en macro-économie (Ioannidis, 2005 ; Chang et Li, 2022), disciplines dans lesquelles une partie des données est soupçonnée d'être des faux positifs.

9. Mais l'incitation à publier ses données en accès ouvert dans le respect des principes FAIR, conduit aussi inévitablement à réformer profondément les pratiques d'évaluation si l'on veut qu'elle soit suivie d'effets. De fait, en l'absence de reconnaissance institutionnelle du travail long et complexe lié à la constitution et au partage des jeux de données, la tendance serait plutôt, du côté des chercheurs, à la fermeture de leurs données (Rebouillat, 2019) afin de pouvoir les exploiter pour leur propre compte. On peut toutefois espérer que les choses soient en train de changer avec l'adoption récente de la déclaration de San Francisco (DORA)[6] par de nombreuses institutions et universités en matière d'évaluation pour la promotion et la carrière des chercheurs. Dès lors que les règles d'évaluation ne s'appuient plus sur les seuls indices bibliométriques liés aux publications dans les revues, comme cela a été le cas depuis des dizaines d'années, mais bien à toute la production scientifique, comme cela est proclamé dans

6. Declaration on Research Assessment : https://sfdora.org/.

la déclaration DORA, on peut alors supposer que le *data paper* a un bel avenir devant lui.

Les origines des *data papers*

10. La première revue entièrement dédiée à la publication de *data papers* semble être le *Journal of Chemical and Engineering Data* de l'*American Chemical Society* créée en 1956. Dans la même veine, la création de catalogues et de guides pour la notation des données a été le premier pas vers la gestion des données de recherche en biologie. Les sommes investies dans la collecte d'échantillons ont incité les chercheurs internationaux à fédérer les données afin que tous puissent y avoir accès sans avoir à investir d'importantes sommes d'argent supplémentaires. La nécessité de rendre les données massivement accessibles a contribué à l'essor des *data papers* en biologie dès les années 1990.

11. Mais les sciences naturelles ne sont pas les seules disciplines à mettre en œuvre des politiques de mise en commun des jeux de données, l'informatique et plus spécifiquement les développeurs s'y consacrent également. Les chercheurs sont en effet également encouragés à partager non seulement les données de la recherche et leurs résultats, mais aussi les méthodes, les outils, les codes source… Partager puis réutiliser le code développé par d'autres est un des principes fondamentaux du développement informatique à l'instar de Git. Créé en 2005, ce logiciel libre de gestion de versions est le système de base du site GitHub, le plus important hébergeur de code informatique au monde où chaque internaute peut récupérer le code mis à disposition par d'autres très facilement. Toutefois, la documentation du code n'a pas toujours été une priorité pour les développeurs avec pour conséquence de rendre difficile la réappropriation et réutilisation des morceaux de code.

12. Afin de pallier ce manque d'information, certains chercheurs du laboratoire Meta AI ont lancé l'initiative *Papers with Code*[7], une ressource libre et ouverte avec des articles sur l'apprentissage automatique du code, des données, des méthodes et des tableaux d'évaluation, tout cela publié avec une licence CC-BY-SA. Tout le monde peut contribuer et proposer une évaluation. D'autres initiatives existent en informatique ; ainsi, dans le domaine du *machine learning*, Gebru *et al.* (2021) proposent que chaque jeu de données soit accompagné d'une fiche technique qui documente sa motivation, sa composition, son processus de collecte et les utilisations recommandées — une démarche proche de celle du *data paper* et des modèles (*templates*) associés aux données décrites dans la publication.

13. En SHS, le *Journal of Digital History* s'est spécialisée dans la production d'articles avec différents *layers* (couches), dont un *data layer*, et une construction éditoriale à plusieurs étapes (Clavert et Fickers, 2022). D'autres revues ont tenté d'intégrer les données au sein des articles ou

7. https://paperswithcode.com/

de les associer aux articles sous forme de fichiers joints à la publication, mais cela ne permettait pas de trouver et d'utiliser ces données indépendamment de la publication auxquelles elles étaient attachées (Candela *et al.*, 2015). Les *data papers* s'inscrivent ainsi dans un processus émergent, interdisciplinaire, non stabilisé avec des formats et des pratiques qui varient selon les disciplines. Dans cet ouvrage, nous nous plaçons en observateurs de cette dynamique, avec nécessairement une image figée d'une réalité arrêtée à un temps T qui ne saurait toutefois être considérée comme une description exhaustive, ni comme la représentation d'une évolution aboutie, mais plutôt comme une invitation au partage.

Le *data paper* et ses enjeux en SHS

14. Le présent ouvrage fait suite aux journées du colloque DH Nord 2021[8] sur les *data papers* et leurs enjeux dans le cadre des humanités numériques. L'objectif ici est avant tout de dégager au milieu des nombreuses initiatives ce qui semble se dessiner pour l'avenir en SHS et d'encourager le lecteur à s'approprier les expériences qui lui sont proposées par nos différents contributeurs, que ce soit en termes de traitement, de rédaction, de partage, d'évaluation, de qualité, de compétences, d'outils et d'infrastructures. Nous avons aussi souhaité que certains chapitres soient suivis d'une discussion visant à éclairer le lecteur sur les aspects techniques et les choix faits par les auteurs.

15. La première partie du livre est consacrée au *data paper* comme une nouvelle forme de publication scientifique en SHS. On assiste depuis plusieurs années en SHS à un « tournant quantitatif »[9] quant au volume de données disponibles sur le web. Avec l'arrivée d'Internet, puis des bases et des catalogues de données en ligne, il existe aujourd'hui un gisement de ressources sans précédent appelé à sans cesse s'enrichir. Tout comme les publications, le partage des données exige d'être structuré et impose de nouveaux modèles d'écriture et de production. Les *data papers* constituent précisément un de ces nouveaux modèles de publication dans les revues à côté des articles de recherche. Mais qu'est-ce qui en fait leur spécificité ? Existe-t-il une définition de référence ? Victoria Le Fourner et Joachim Schöpfel[10] se proposent dans une description détaillée des pratiques d'en définir les contours. Si la question de la définition du *data paper* s'attache de prime abord à l'article en soi, elle est indissociable de celle de la donnée, mais aussi de la métadonnée.

16. La citation est le point d'orgue de toute publication scientifique et le *data paper* n'en est pas exempt. Au contraire, elle peut avoir un impact encore plus important que pour les articles de recherche puisqu'elle concerne simultanément l'entrepôt de données, le producteur de données, la revue, l'auteur et enfin le réutilisateur. Sur la base d'un formulaire d'enquête à l'adresse des chercheurs en

8. DH Nord 2021 https://dhnord2021.sciencesconf.org/.

9. Voir le chapitre de Victor Gay dans cet ouvrage : « Un *data paper* en SHS : pourquoi, pour qui, comment ? ».

10. Voir le chapitre de Victoria Le Fourner et Joachim Schöpfel dans cet ouvrage : « Le paysage des *data papers* ».

archéologie, Violaine Rebouillat[11] nous livre une analyse de leurs pratiques de citations des *data papers*, ainsi que de leurs motivations. C'est aussi l'occasion de se demander en quoi le *data paper* modifie les comportements par rapport à la citation du jeu de données dans un entrepôt. Les revues qui publient des *data papers* sont bien entendu à la charnière des pratiques de structuration, de publication, de citation et de réutilisation des données. Les pionnières dans ce domaine, comme le *Journal of Open Humanities Data*[12] nous exposent les difficultés notamment liées à l'hétérogénéité des disciplines en SHS tant du point de vue des pratiques de partage que de la nature même des « données ».

Écrire des *data papers* en SHS

17. La deuxième partie du livre, intitulé « Écrire des *data papers* en SHS », contient plusieurs exemples et partages d'expériences. Le premier chapitre[13], par Victor Gay, insiste sur le fait que, dans tous les cas, quelle que soit leur forme, les *data papers* ont en commun d'ouvrir l'accès aux données de toutes natures produites par le chercheur et d'accompagner dans leur réutilisation à partir d'une publication validée par les pairs. La contribution de Mareike König et ses collègues[14] nous apporte un exemple concret de pratique du *data paper* en histoire dans lequel la structuration de l'article est laissée au libre choix des auteurs. Ainsi, l'*Adressbuch 1854* se présente dans une forme qu'on pourrait considérer comme hybride, avec une première partie proche d'un article de recherche et une seconde partie dédiée à la description des données. Ce format qui permet de restituer dans sa totalité le projet, tant épistémologiquement que techniquement, serait plus adapté au sujet traité et à la discipline historique.

18. Un autre exemple nous est apporté avec la publication d'un article exécutable[15] qui met également en lumière les problématiques liées à la crise de reproductibilité et de réplicabilité que le *data paper* contribue à résoudre. La mise à disposition sous la forme d'un *Notebook* de l'ensemble du protocole, des données et du code informatique sur le traitement automatique des langues et de son usage pour la classification automatique des travaux universitaires montre à quel point le partage des données et des codes sources ne se limite pas à une discipline, mais *a contrario* stimule le partage interdisciplinaire.

19. Toutes les revues ne proposent pas pour l'instant de modèle et d'instructions standardisés pour la rédaction

11. Voir le chapitre de Violaine Rebouillat dans cet ouvrage : « Révéler les formes et logiques de citation des *data papers* en archéologie : le cas du *Journal of Open Archaeology Data* ».
12. Voir le chapitre de Paola Marongiu, Nilo Pedrazzini, Marton Ribary et Barbara McGillivray dans cet ouvrage : « Le *Journal of Open Humanities Data* (*JOHD*) : enjeux et défis dans la publication de *data papers* pour les sciences humaines et sociales (SHS) ».
13. Voir le chapitre de Victor Gay dans cet ouvrage : « Un *data paper* en SHS : pourquoi, pour qui, comment ? »
14. Voir le chapitre de Mareike König, Gérald Kembellec et Evan Virevialle dans cet ouvrage : « *Data paper* en humanités numériques : Adressbuch 1854 ».
15. Voir le chapitre de Vincent Arnaud, Kevin Bouchard et Gilles-Philippe Morin dans cet ouvrage : « Utiliser un *data paper* en traitement automatique des langues : un exemple de classification automatique de mémoires et de thèses universitaires ».

des *data papers*, avec pour conséquence l'absence constatée[16] de certaines données contextuelles pertinentes dans la publication, qui peut pénaliser la possible réutilisation des données par les chercheurs. Cette difficulté pose en effet problème en SHS devant l'hétérogénéité des données et des pratiques disciplinaires. Dans ce cas, l'évaluation par les pairs peut jouer un rôle déterminant pour faciliter la reproductibilité.

Qualité, évaluation, compétences

20. Aussi, la troisième partie du livre intitulé « Qualité, évaluation, compétences » fait le lien entre l'évaluation et la qualité des *data papers* et les compétences mises en œuvre, avec au centre, l'enjeu des principes FAIR. La mise en place par la revue *Cybergeo* de critères spécifiques d'évaluation des *data papers*[17] vise précisément à normaliser la présentation de ces nouveaux articles tout en les accompagnant d'un *template* adapté à la thématique traitée. Cette normalisation s'accompagne toutefois du travail de FAIRisation des données[18], un processus destiné à répondre aux objectifs de réutilisation. La question des droits d'auteurs est au cœur des principes FAIR et nombre de licences de libre accès, telles que les Creative Commons, s'emploient à y apporter une réponse, toutefois, là encore, la très grande variété des définitions sur la « donnée » ne permet pas de garantir une couverture juridique adaptée. Une nouvelle approche du droit de l'UE en matière de réglementation de l'accès aux données et de leur réutilisation dans le cadre de la science ouverte et plus largement est en cours[19].

21. Le Plan de Gestion de Données (PGD) est un document qui est désormais exigé pour tous les projets financés sur fonds publics. Certains services communs de documentation des universités proposent de former et accompagner à la rédaction de ce document évolutif. On ne peut s'empêcher de voir dans le PGD des similitudes avec le *data paper* notamment dans la documentation et l'application des principes FAIR. Dès lors de s'interroger, dans quelle mesure ce dispositif peut-il faciliter la préparation d'un *data paper* alors que le travail de conformité des données a déjà eu lieu en amont[20] ?

22. Le dernier chapitre décrit la structuration et la distribution de corpus de langage oral et montre les similitudes de leur diffusion avec celle des *data papers*. Dans ce chapitre, Christophe Parisse[21] décrit également comment passer d'un corpus à un *data paper*, avec les avantages pour la recherche dans le domaine du langage oral.

16. Voir le chapitre de Jihyun Kim dans cet ouvrage : « Une analyse des modèles et instructions des *data papers* : types d'informations contextuelles décrites par les data journals ».
17. Voir le chapitre Clémentine Cottineau-Mugadza, Christine Kosmopoulos et Denise Pumain dans cet ouvrage : « Évaluer un *data paper*, l'exemple de *Cybergeo* ».
18. Voir le chapitre d'Alain Rivet dans cet ouvrage : « La FAIRisation des données ».
19. Voir le chapitre de Thomas Margoni, Luca Schirru et Brad Spitz dans cet ouvrage : « Le rôle des licences dans la FAIRisation des données ».
20. Voir le chapitre d'Alicia León y Barella dans cet ouvrage : « Science ouverte, plans de gestion de données et data papers au cœur d'une offre de services : l'exemple du SCD de l'université de Lille ».
21. Voir le chapitre de Christophe Parisse dans cet ouvrage : « Des corpus de langage oral aux *data papers* ».

Outils et infrastructures

23. La quatrième partie du livre présente quelques outils et infrastructures dans l'écosystème des *data papers*, dont notamment le nouveau dispositif Recherche Data Gouv progressivement mis en place par le ministère de l'Enseignement supérieur et la Recherche depuis 2022 ; le premier chapitre décrit l'intérêt de ce dispositif pour la production et la diffusion des *data papers*[22]. La convergence dans les pratiques de traitement des données et de leur partage inspire d'autres initiatives du côté des infrastructures, comme Huma-Num, qui consistent à développer des interactions entre les différents services liés à la donnée afin de les regrouper dans un nouvel écosystème plus fonctionnel[23].

24. Alix Chagué et ses collègues décrivent l'initiative pour une mutualisation des données issues de la reconnaissance des écritures manuscrites (HTR) susceptible d'offrir un cadre pour la construction de *data papers* standardisés dans ce domaine[24]. Le chapitre suivant restitue la mise en œuvre d'un entrepôt Dataverse au sein des laboratoires de Sciences Po Paris (*data.sciencespo.fr*) et pose la question des *data papers* dans un tel contexte, en lien notamment avec les enjeux de la documentation des données et de l'interconnexion avec les revues[25]. Cette interconnexion entre données et publications est au centre également du dernier chapitre[26], mais dans un tout autre domaine, celui de la biodiversité. Et cette fois-ci, il s'agit d'une expérimentation d'une rédaction assistée d'un *data paper* à partir des métadonnées d'un entrepôt de données, ce qui nécessite un haut degré d'interopérabilité des systèmes et de standardisation des métadonnées.

25. Dans un dernier chapitre[27], nous présentons quelques perspectives pour le futur développement de ce nouveau type de publication scientifique qu'est le *data paper*. Ce chapitre revient notamment sur le rôle des *data papers* dans l'écosystème émergent des infrastructures des données de recherche, sur la génération automatique d'un *data paper*, sur l'importance des métadonnées, sur le lien avec les principes FAIR et sur plusieurs variantes proches du *data paper*. Pour finir, le chapitre (re)pose la question de l'intérêt de ce nouveau type de publication pour les SHS.

22. Voir le chapitre de Christine Kosmopoulos et Joachim Schöpfel dans cet ouvrage : « Les *data papers* dans l'écosystème Recherche Data Gouv ».

23. Voir le chapitre de Nicolas Sauret, Stéphane Pouyllau et Mélanie Bunel dans cet ouvrage : « Vers un écosystème d'écriture et d'édition avec les données ».

24. Voir le chapitre d'Alix Chagué, Thibault Clérice et Laurent Romary dans cet ouvrage : « HTR-United : un écosystème pour une approche mutualisée de la transcription automatique des écritures manuscrites ».

25. Voir le chapitre d'Alina Danciu, Anna Egea, Guillaume Garcia et Cyril Heude dans cet ouvrage : « De l'entrepôt de données aux data papers. Retour sur l'expérience de Data Sciences Po ».

26. Voir le chapitre de Sophie Pamerlon dans cet ouvrage : « Le data paper appliqué à la biodiversité : standards, outils et processus mis en œuvre pour démocratiser le concept dans la communauté de la bio-informatique ».

27. Voir le chapitre de Christine Kosmopoulos, Victoria Le Fourner et Joachim Schöpfel dans cet ouvrage : « Perspectives ».

Le *data paper*, une nouvelle forme de publication scientifique en SHS

Le paysage des *data papers*

Victoria Le Fourner et Joachim Schöpfel

Concept et contexte

Des définitions

1. De manière laconique, Callaghan *et al.* (2012) décrivent le *data paper* comme un document contenant des informations sur le quoi, le où, le pourquoi, le comment et le qui des données. Encore plus synthétique, le *Web of Science* indique qu'un *data paper* fournit des « faits sur les données ». Une définition plus complète est proposée par l'*International Society for Knowledge Organization* (ISKO) qui définit les *data papers* comme des articles rédigés, révisés par des pairs et citables, publiés dans des revues scientifiques, dont le contenu principal est une description des données de recherche publiées, ainsi que des renseignements contextuels sur la production et l'acquisition de ces données, dans le but de faciliter la « trouvabilité » et la réutilisation des données de recherche ; ils sont intégrés à la gestion des données de recherche et liés à des entrepôts de données (Schöpfel *et al.*, 2020).

2. Les *data papers* permettent à la fois de trouver les données, mais informent aussi sur leur provenance, sur leur intérêt potentiel et sur les droits de leur réutilisation. De plus, le fait d'être publié dans une revue scientifique implique normalement l'évaluation par les pairs et l'attribution d'un identifiant pérenne, comme le *Digital Object Identifier* (DOI).

3. Dans certaines revues, les données sont ajoutées aux articles sous forme de matériel supplémentaire ; ailleurs, elles font l'objet d'un article particulier (Kratz et Strasser, 2015 ; L'Hostis *et al.*, 2016). Si la description des données n'est pas publiée sous forme d'un article de revue, il ne s'agit pas d'un *data paper*. Aussi, l'encyclopédie internationale de la gestion des connaissances de l'ISKO préfère le terme de *data document*, plus large et inclusif que celui de *data paper*[1].

4. La question d'une définition du *data paper* renvoie à avoir un champ défini et circonscrit de ce qu'est une donnée de recherche (*research data*), de ses standards et des tenants et aboutissants de sa publication. Beaucoup a été dit et écrit sur les données de recherche sans qu'il y ait, là non plus, de définition précise et consensuelle. Citons à titre d'exemple celle de l'Organisation de coopération et de développement économiques (OCDE) selon laquelle les données de recherche sont « des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés

1. Encyclopedia of Knowledge Organization https://www.isko.org/cyclo/data_documents.

comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche » (OCDE, 2007). Selon Borgman, il serait toutefois plus réaliste de comprendre les données de recherche comme un « objet-frontière », c'est-à-dire un objet ou arrangement qui permet à différents groupes de travailler ensemble sans consensus préalable (Star, 2010) ; en principe, tout peut être ou peut devenir données, et les enquêtes sur les données de recherche révèlent que c'est surtout le regard du chercheur qui donne sens et valeur à la donnée, dans la mesure où elle est construite par l'observateur (Borgman, 2015).

5. À ceci s'ajoutent en SHS deux particularités : l'importance de la dimension interprétative pour l'encodage des données, lors de l'écriture des métadonnées descriptives et pour les annotations, et l'importance de la contextualisation des données en question, c'est-à-dire la connaissance de leur acquisition, production ou construction, de la contribution des personnes à l'origine des données, etc. (Schöpfel, 2020). On voit ici le rôle primordial que jouent les métadonnées pour la réutilisation des données de recherche. Aussi, comme certains auteurs les décrivent comme carburant pour l'économie (Neuroth *et al.*, 2013), on pourrait assimiler les *data papers* à une sorte de pipeline pour ce carburant. Cependant, les *data papers* ne diffusent pas seulement les métadonnées, ils y ajoutent aussi de la valeur, par le biais d'informations contextuelles plus riches, d'évaluations, de nouveaux identifiants, etc.

La différence avec les articles scientifiques

6. La plupart de ces définitions ont en commun qu'elles insistent sur la différence avec les articles scientifiques « classiques » qui, pour reprendre les mots du Plan national pour la science ouverte, « exploitent, analysent et interprètent les données scientifiques », tandis qu'un *data paper* se contente de décrire « finement un/des jeu(x) de données de façon à en faciliter la compréhension et l'éventuelle réutilisation ». La même distinction est faite par le *Web of Science* : un *data paper* ne fournit pas d'analyse ou de recherche à partir des données, contrairement aux articles classiques. Plus simplement, les *data papers* ne portent pas sur la recherche, mais sur des données (Smith, 2012) ; ils ne se définissent pas par la présence de telle ou telle information, mais par l'absence d'analyse ou de conclusions.

7. Pas d'hypothèse, pas d'analyse, pas d'interprétation, pas de conclusion. La distinction cruciale avec l'article scientifique semble donc se faire sur l'absence de résultats, de discussion et de conclusion. La revue *Data in Brief* d'Elsevier recommande à ce titre aux auteurs d'éviter d'utiliser des mots tels que « résultats », « études » ou « conclusions ». Comparées aux articles scientifiques traditionnels, les informations ne sont pas organisées de la même manière. En théorie, articles de recherche et articles de données peuvent avoir certaines parties en commun (introduction, contexte, méthodologie) tandis que d'autres sections manquent aux articles de données (résultats, discussion, conclusion).

8. La réalité, cependant, est plus nuancée, et la distinction entre ces deux types d'articles n'est pas aussi stricte que cela (Li *et al.*, 2020). Une raison est la diversité des articles scientifiques dont les structures et les contenus restent assez variables, malgré un format largement accepté et employé surtout en sciences expérimentales (IMRaD : introduction, méthodologie, résultats et discussion). Une autre raison est qu'ils partagent la procédure de sélection (évaluation par les pairs, acceptation par la rédaction) et la « citabilité » dans une revue, avec l'attribution d'un identifiant pérenne (le plus souvent, un DOI de CrossRef). Mais, comme nous allons voir plus loin, la raison principale est que les *data papers* eux-mêmes ne se limitent pas toujours à la simple description des données. Loin de faire une distinction nette, l'analyse sociolinguistique révèle une identité à double face et entrelacée des *data papers* (Li et Jiao, 2022).

9. Un *data paper* peut-il remplacer un article de recherche ? La réponse paraît simple : non, bien sûr, car il manquerait l'analyse et la discussion des résultats. De l'autre côté, le développement rapide du paradigme de la *data-intensive scientific discovery* (Hey *et al.*, 2009) interroge l'intérêt et la pertinence même des articles de recherche classiques. La question reste donc ouverte, comme celle de la complémentarité des deux types d'articles.

10. Faut-il d'abord publier un article « normal », puis un *data paper* pour valoriser le partage des données ? Ou dans l'autre sens ? Le plus souvent, c'est d'abord l'article de recherche (Thelwall, 2020). Dans le cadre de l'ERC *Desert Networks* (2018-2022), par exemple, la démarche de publication se fait en trois temps, d'abord avec l'article principal, publié dans le *Journal of Computer Applications in Archaeology* où les auteurs présentent la démarche de modélisation des itinéraires entre les sites archéologiques et les résultats. Dans un second temps, un *data paper* est publié dans le *Journal of Archaeology Data* qui présente le jeu de données, sa construction et sa qualité ; et enfin le dépôt est fait dans Zenodo : données d'entrées, résultats, métadonnées et outils de traitement[2].

11. De l'autre côté, la revue *Cybergeo* annonce que ses *data papers* sont des articles scientifiques autonomes dont la publication peut être suivie d'un autre article de recherche exploitant l'analyse de ces données géographiques. Et puis, autre question, faut-il pour chaque article de recherche un *data paper ?* Ou un *data paper* pour l'ensemble des données d'un projet, même s'il y a plusieurs articles de recherche ? En fait, il n'y a pas de réponse unique à toutes ces questions. Dans les faits, c'est la pratique au sein d'une communauté de données qui fournira la réponse – dans un domaine particulier, autour d'un équipement scientifique, avec une méthodologie spécifique, affiliée à une institution de recherche, liée à une revue ou une plateforme, etc. Parfois, il peut s'agir tout simplement d'un choix décidé par une équipe de recherche.

2. Cf. Stéphane Renault (CNRS) lors de l'*Atelier Données de la MITI* (CNRS). Réseau Médici (5 novembre 2020). Introduction du webinaire. [Vidéo]. Canal-U. https://www.canal-u.tv/81841.

12. Par ailleurs, Candela *et al.* (2015) font remarquer que les éditeurs traitent les *data papers* exactement comme des articles scientifiques par rapport aux droits d'auteur (copyright) et par rapport aux licences de diffusion (Creative Commons) ; il n'y a pas de différences de ce côté-ci.

Le nombre

13. Peut-on faire une estimation du nombre de *data papers* ? Le *Web of Science* indexe les *data papers* depuis quelques années et en recense en mars 2023 presque 12 000, dont une grande partie publiée depuis 2018 (figure 1).

14. Après une croissance rapide à partir de 2016, leur nombre semble se stabiliser autour de 2 000 articles par an, sans progression significative. C'est peu, comparé à la totalité des articles scientifiques, moins de 0,1 % de la production annuelle. Mais le *Web of Science* ne couvre qu'une partie de la production scientifique, et le nombre réel des *data papers* est sans doute supérieur. Avec *InCites* par exemple, Adamczak (2021) a réussi à identifier plus de 15 000 articles de données, publiés entre 1980 et 2020 majoritairement par des auteurs aux États-Unis, en Chine, au Royaume-Uni, en Italie et en France, dont beaucoup de chercheurs affiliés au CNRS, à l'INRAE et à l'IRD.

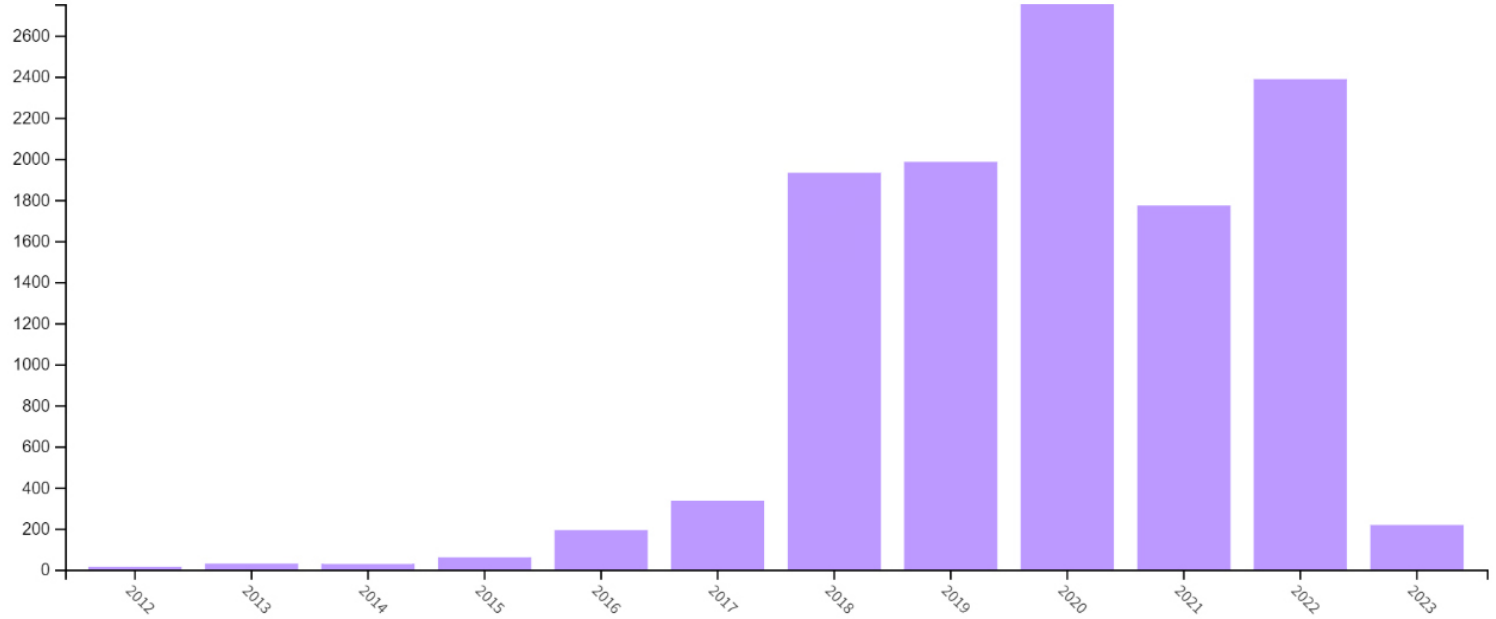


Figure 1. Le développement du nombre de data papers 2012-2022.
Crédit : *Web of Science*, 21 mars 2023.

Les revues de données

15. Environ trois quarts de ces *data papers* sont publiés par seulement deux méga-revues multidisciplinaires, *Data in Brief* d'Elsevier et *Scientific Data* de Springer Nature (Thelwall, 2020). L'analyse du *Web of Science* confirme ce chiffre (figure 2).

Figure 2. Nombre de *data papers* par revue (N =11 701 *data papers*)
Crédit : *Web of Science*, 21 mars 2023.

16. Quelques revues – environ 30 en 2019 – sont entièrement dédiées à la publication de ces *data papers*. Un nombre plus important – au moins 50 – publient des *data papers* avec d'autres types d'articles. Leur nombre réel est sans doute plus élevé, car ces revues n'indiquent pas toujours explicitement qu'elles publient des *data papers* (absence de rubrique…), ou elles les appellent autrement. Ces revues publient souvent entre 10 et 100 *data papers* par an et couvrent surtout des domaines STM.

17. Cependant, il n'existe pas de répertoire pour ces revues, et le référencement dans les bases de données n'est pas optimal (Li *et al.*, 2021). Ainsi, une autre étude récente (Walters, 2020) identifie 128 revues avec des *data papers*, dont 19 revues de données « pures » (pour lesquelles les articles de données constituent au moins la moitié des articles de la revue) et 109 autres revues qui ont commencé à publier des *data papers* au même titre que les articles de recherche classiques, avec parfois une rubrique dédiée ou sous forme de *varia*, mais dans lesquelles ces articles ne représentent en moyenne que 1,6 % de tous les articles. Un exemple récent en SHS de cette deuxième catégorie est la *Revue française des Sciences de l'Information et de la Communication* (RFSIC).

18. D'après Adamczak (2021), 69 % des revues de données contiennent des *data papers* qui relèvent des sciences naturelles, 30 % de la médecine et 17 % des sciences de l'ingénieur et de la technologie. Les premières revues de ce genre ont été lancées dans les années 50 et 60, mais

la plupart sont beaucoup plus jeunes (Thelwall, 2020). Aujourd'hui, tous les grands éditeurs scientifiques ont lancé leurs *data journals* – plus de la moitié appartiennent aux cinq éditeurs commerciaux les plus importants (Elsevier, Springer Nature avec Biomed Central, Wiley, Taylor & Francis, SAGE) (figure 3).

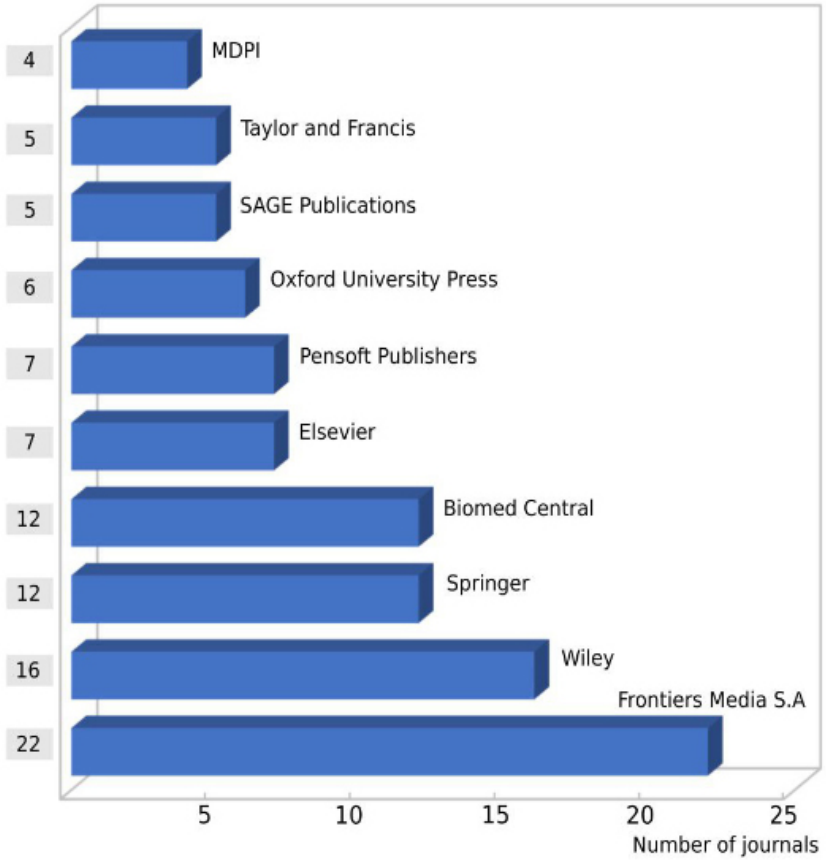


Figure 3. Le nombre des revues de données par éditeur (N =96)
Crédit : Adamczak, 2021.

19. D'après la même étude, 19 % des revues de données publient des *data papers* en SHS (Adamczak, 2021). Toutefois, leur nombre absolu est bas. Dans la base de données *Web of Science*, moins de 1 % des *data papers* relèvent des SHS – il y en a en archéologie et en gestion, mais aussi un peu en psychologie, en sciences de l'éducation et en sociologie. Ces chiffres sont sous-estimés, car le *Web of Science* ne couvre pas bien les SHS ; mais il est peu probable que la part réelle des *data papers* en SHS soit beaucoup plus importante, en comparaison notamment avec les sciences de la vie et la médecine.

20. On trouve deux revues de données en archéologie et deux revues en SHS, le *Journal of Open Humanities Data Research Data* chez Ubiquity Press (racheté par De Gruyter en 2022) et le *Research Data Journal for the Humanities and Social Sciences* chez Brill. Néanmoins, la revue la plus importante pour les *data articles* en SHS est de loin la revue multidisciplinaire d'Elsevier, *Data in Brief* dont environ 10 % d'articles relèvent des SHS ; qui plus est, ces articles correspondent à environ deux tiers de tous les *data papers* en SHS.

21. Parmi ces revues d'articles, le libre accès avec *article processing charges* (APC) est le modèle d'affaires dominant. La grande majorité des *data papers* est donc librement et gratuitement accessible, souvent avec une licence Creative Commons. Concrètement, des 28 revues d'une étude de 2019 (Schöpfel *et al.*, 2019) qui publient uniquement (ou majoritairement) des *data papers*, 19 correspondaient à ce modèle (revue en libre accès avec APC), 4 étaient des

revues hybrides (revues sur abonnement, avec des articles en libre accès) et 2 étaient des revues « diamant » (libre accès sans APC). L'étude de Candela *et al.* (2015) aboutit au même résultat : de 116 revues (avec et sans autres types d'articles), seulement 3 ne sont pas en accès libre. Quant au *Web of Science*, 97 % des *data papers* sont en libre accès.

Fonction et contenu

22. Pour certains auteurs, écrire un *data paper* revient à compléter les parties vides, un peu comme composer un sonnet (Kembellec et Le Deuff, 2022). Ainsi, Nilo Pedrazzini a pu tweeter au moment du colloque DHNord 2021 à Lille[3] que « si nous avons les données et que nous pensons qu'elles ont une certaine valeur, la seule chose à faire est de suivre la structure et de remplir le vide, le squelette ». La structure est imposée, rigide et elle ne nécessite pas une réflexion spécifique ; en revanche, le contenu et la place accordée à celui-ci varient beaucoup.

23. D'une manière générale, les *data papers* sont souvent plus courts qu'un article « normal », mais peu de revues exigent explicitement des articles courts, limités par exemple à 3 000 mots maximum. On peut ainsi trouver des *data papers* aussi longs qu'un article de recherche, voire plus longs, avec des dizaines de pages.

3. *#dhnord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*. https://www.meshs.fr/page/dhnord2021.

24. Ce qu'il faut toutefois bien garder à l'esprit à la lecture d'un *data paper*, c'est qu'il ne s'agit pas nécessairement des données de tout un projet. En effet, ils présentent souvent un ensemble de données unique et clos, et peu de *data papers* décrivent plusieurs jeux de données ou une base de données complète.

La fonction

25. Plus haut, nous avons décrit la fonction essentielle d'un *data paper* qui est (pour reprendre les mots de Callaghan *et al.*, 2012) censé contenir des informations sur « le quoi, le où, le pourquoi, le comment et le qui des données », dans le but de faciliter la « trouvabilité » et la réutilisation des données. Un *data paper* remplit donc sa mission dans la mesure où il contribue à identifier et à trouver des données, à l'aide d'identifiants et de liens, d'une description plus ou moins riche, complémentaires aux métadonnées des données et d'une information sur les conditions d'accès et de réutilisation. C'est une fonction de facilitation, en quelque sorte.

26. Le *data paper* fait partie des bonnes pratiques de la gestion des données de recherche, tout comme les différentes versions du plan de gestion ou, dans certains cas, la partie « traitement de données » du dossier éthique d'un projet de recherche. En fait, il n'est pas si loin de la version finale d'un plan de gestion, c'est-à-dire de la mise à jour définitive du plan après la fin d'un projet de recherche, avec des informations sur le traitement des données pendant et après

la fin du projet, sur la nature des données, sur la méthodologie et les normes appliquées, et sur les modalités de préservation, de partage et d'accès[4].

27. Un autre élément fonctionnel et caractéristique est le lien avec les entrepôts de données. Le *data paper* joue son rôle uniquement dans la mesure où il renvoie vers des données déposées et accessibles dans une infrastructure de recherche, c'est-à-dire en règle générale, dans un des nombreux entrepôts de données. Sans ce lien, le *data paper* perd son intérêt. Pour le dire autrement, il n'a pas de valeur intrinsèque.

28. Reste une question : faciliter la recherche et la réutilisation, oui, mais pour qui ? La réponse est double : d'abord et avant tout, pour les chercheurs, professionnels et d'autres personnes (prestataires de services, développeurs, industriels, journalistes, etc.) potentiellement intéressées par ces données ; mais la mission des *data papers* ne s'arrête pas là, car elle inclut également les machines (moteurs de recherche, agrégateurs et entrepôts de données, etc.), d'où une certaine rigidité du format et un haut degré de standardisation d'une bonne partie de *data papers*.

4. Cf. par exemple, les critères des plans de gestion des projets européens https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.

La structure

29. L'analyse des instructions aux auteurs et des modèles d'article (*templates*) révèle quelques éléments constitutifs et qu'on retrouve systématiquement dans les *data papers*. D'après Kim (2020), un *data paper* est composé de quatre sections : les caractéristiques générales des données, suivies des informations sur leur production, sur leur dépôt et sur leur réutilisation. L'analyse de Schöpfel *et al.* (2019) montre une plus grande diversité (figure 4).

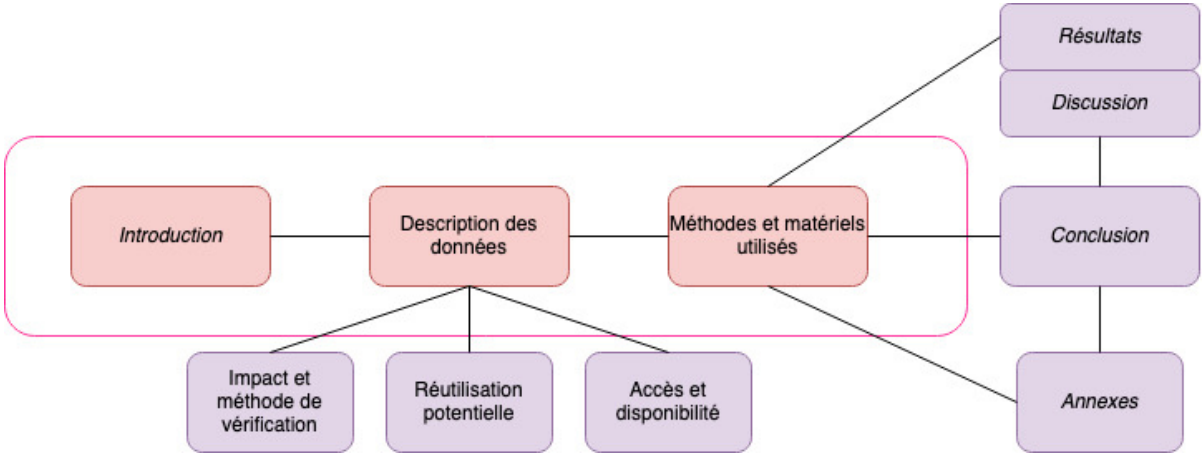


Figure 4. La structure d'un *data paper* (d'après Schöpfel *et al.*, 2019).

30. Trois sections constituent le « cœur » d'un *data paper* :

1. L'introduction résume le projet de recherche (contexte) dont sont issues les données. Au sein de cette première partie, les auteurs sont amenés à expliquer les travaux antérieurs sur le sujet lié aux données, fournir des informations sur l'initiative qui a conduit à la production de ce jeu de données, quels contributeurs

et quelles sources de financement (Kim, 2020). Cette partie n'est généralement pas la plus longue ; elle pose le cadre.

2. Une autre section décrit les données d'une façon plus ou moins standardisée, à partir des métadonnées dans un format générique (multidisciplinaire) ou plus spécifique à un domaine ou un équipement. Cette description est destinée à des usagers extérieurs au projet, pas forcément familiers du milieu de production des données et des choix de classement par exemple. Une enquête sur la réutilisation révèle que personne ne parvient à utiliser les données « telles quelles » et qu'un échange avec le producteur de données est toujours nécessaire (Kunze *et al.*, 2011). Les parties description des données et méthodes doivent correspondre à cet échange entre chercheurs. Au sein de cette partie, il est utile d'expliquer l'arborescence des fichiers et les formats utilisés afin que l'usager puisse prendre en main le plus aisément possible les données et se repérer dans des fichiers qui ne sont pas les siens. Dans cette veine, *Cybergeo* recommande aux auteurs d'articles sur des bases de données géographiques de « préciser les échelles concernées, les composantes spatiales, la géométrie et la compatibilité avec les systèmes d'information géographique des éléments inclus dans la base de données, en plus des informations requises habituellement »[5].
3. Une troisième section décrit la méthodologie et les équipements, instruments, etc., à l'origine de la collecte et production des données. Cette partie correspond à la description du protocole de production des données, des méthodes et/ou du matériel utilisés. Les auteurs y retracent les traitements et les procédures mis en place en plusieurs points afin de permettre aux lecteurs de saisir les étapes par lesquelles sont passées les données. Le lecteur doit être informé à travers cette partie de l'histoire des données afin de ne pas se méprendre sur ses origines et comprendre les potentiels biais qu'il pourrait y avoir par rapport à la réutilisation envisagée. Certains auteurs ajoutent directement les questions éthiques, tandis que d'autres préfèrent consacrer une petite partie à la fin du *data paper* pour les préciser, dans le cadre des potentielles réutilisations. Certaines revues recommandent d'expliquer les traitements, par exemple en définissant les paramètres spécifiques, d'une manière détaillée et compréhensible pour des personnes extérieures aux champs disciplinaires.

31. Trois autres rubriques sont caractéristiques pour les *data papers*. Dans certaines revues, elles font partie de la description des données tandis qu'ailleurs, elles constituent des sections à part entière :

1. Une section sur la valeur des données, en particulier sur les méthodes de contrôle de qualité et de validation des données.
2. Une section sur l'intérêt potentiel de ces données pour d'autres chercheurs, avec par exemple des cas de réutilisation, des traitements possibles, etc.
3. Une troisième section sur l'endroit où se trouvent les données (entrepôts…) et sur les droits d'accès et de réutilisation (licences, restrictions, etc.). Il est rarement demandé aux auteurs de fournir des informations sur

5. « Présentation de la rubrique Data Papers », *Cybergeo : European Journal of Geography*. https://journals.openedition.org/cybergeo/28545.

la réputation du dépôt ou sur les pratiques et garanties de conservation.

32. Enfin, on trouve plus ou moins régulièrement d'autres sections qui ne sont pas spécifiques aux *data papers*, dans la mesure où on les trouve également dans d'autres types d'articles – une conclusion, des annexes, et parfois même des résultats et une discussion. De même, comme dans n'importe quel article de recherche, on retrouve des déclarations d'éventuels conflits d'intérêts et/ou d'éthique, des remerciements, une description de la contribution de chaque auteur (micro-attribution), etc.

Accès et réutilisation

33. Guidé par le mode d'emploi que constitue le *data paper*, le lecteur, en suivant les conditions des licences adoptées, devrait être en mesure de reproduire l'analyse et arriver aux mêmes résultats. La reproductibilité est au cœur des enjeux du *data paper*. Souvent scindés en plusieurs courtes parties indépendantes, le contexte actuel des données, ou les liens vers l'endroit de stockage et les conditions d'accès sont donnés à la fin du *data paper*.

34. Ces dernières parties sur la réutilisation des données se composent généralement de quatre catégories : les limites ou anomalies sur le jeu de données, le type ou format de données ainsi que la version. Dans cette dynamique, Gebru *et al.* (2021) recommandent de bien préciser la version du jeu de données et des formats de données ou d'encodage. En effet, d'une version à l'autre de logiciel de traitement, l'analyse n'est pas forcément reproductible. La date de dépôt des données doit aussi être précisée, les conditions d'utilisation de l'entrepôt peuvent changer entre le dépôt et la réutilisation, par une tierce personne, des données. Il est également recommandé aux auteurs de détailler le volume des données et la/les licences utilisées. La licence utilisée permet dès le début à l'utilisateur potentiel de connaître les conditions de la réutilisation des données. Il faut finalement, dans cette dernière partie, donner le plus d'informations susceptibles d'intéresser une personne souhaitant réutiliser les données sans être au courant de son contexte de production. Il est ainsi bienvenu de préciser la langue des données afin de prévoir un public plus large de réutilisateurs.

35. Parfois, un encart en tête ou en fin de l'article récapitule les informations nécessaires, comme le lien direct vers les données (grâce à un DOI de DataCite, le plus souvent), la couverture spatiale et temporelle des données, la licence utilisée, etc. On peut aussi trouver l'information sur l'existence d'un plan de gestion de données (*data management plan*). Comme nous l'avons évoqué plus haut, le but premier d'un *data paper* est de fournir la réponse au qui, quoi, comment, où, pourquoi des données, et si toutes ces réponses se trouvent dans un article, alors, *a priori*, il s'agit bien d'un *data paper*.

Processus

36. Si la définition du *data paper* et de ce qu'il doit contenir fait plutôt consensus, il semblerait qu'il y ait des variantes au sein d'une grande catégorie que seraient les articles sur les données. Le concept de données diverge selon les champs disciplinaires et les possibles traitements. À cet égard, Hisen explique en 2014[6] qu'il faudrait distinguer les articles en fonction du type de données dont il est question. Il propose ainsi trois catégories de données en fonction de leur place dans la chaîne de traitement : les données brutes (issues de l'observation directe, de sondages…), les données lisibles par des machines (logiciels, données stockées électroniquement) et les données issues d'un calcul ou traitement. Ce sont ces dernières qui sont reproductibles à la différence des données brutes qui doivent être mises à disposition pour permettre le calcul.

37. Ces observations se retrouvent dans l'émergence de différentes catégories de *data paper*, qui prennent parfois d'autres noms. Écrire un *data paper* ne signifie donc pas toujours la même chose pour un auteur en fonction de l'étape où la description des données intervient dans la chaîne de traitement et de la spécialité du chercheur. Aussi, cette multiplicité de réalités derrière le nom de *data paper* ou la prolifération de noms dérivés de cette idée risque d'embrouiller les utilisateurs et de rendre l'accès aux données et aux informations sur les données plus complexes dans les revues (Candela *et al.*, 2015).

La motivation

38. Pour comprendre qui sont les auteurs de ce nouveau type de publication scientifique, il est important de prendre en compte les motivations qui poussent les uns et les autres à rédiger un *data paper*. Outre l'idée d'origine de fournir aux auteurs un moyen d'obtenir une récompense via la citation (Chavan et Penev, 2011 ; Parsons et Fox, 2013), d'autres motivations ont été décelées, entre intérêts personnels et intérêt disciplinaire (Lee et Kim, 2021 ; Huang et Jeng, 2022). Une des motivations les plus frappantes est celle de marquer la propriété d'un territoire scientifique (*claim a territory*). Certains auteurs se serviraient ainsi de l'étape du *data paper*, avant de rédiger un article traditionnel, pour faire connaître leurs travaux sur certaines données et revendiquer la primauté de l'origine. Pour certains, les *data papers* seraient une manière d'exprimer leur fierté et de la partager à la communauté (Callaghan, 2013).

39. Cette fierté semble toutefois être à double tranchant. Lors du colloque DHNord 2021, la crainte a été exprimée d'exposer des méthodes ou des données erronées à la communauté scientifique et d'avoir ensuite des retombées négatives touchant sa réputation. Le lien intrinsèque entre le *data paper* et le jeu de données pose donc question pour les auteurs. Ainsi, dans les motivations pour le choix de la revue, les chercheurs regardent les conditions posées par celle-ci sur l'entrepôt. La volonté des éditeurs conjuguée aux dispositions politiques de la

6. Commentaire sur l'article de Kratz et Strasser (2015).

recherche ne rencontre donc pas toujours les motivations des chercheurs.

40. Certains chercheurs utilisent le *data paper* comme double moyen, à la fois pour mettre en valeur le travail réalisé, mais aussi pour structurer une discipline (Knauf, 2022). Cette collaboration des chercheurs avec d'autres collègues, notamment les ingénieurs, est fréquente. Celle-ci peut intervenir directement dans la rédaction du *data paper* ou de manière indirecte avec une aide apportée sur le dépôt de données et la rédaction. On trouve ainsi une augmentation des formations sur l'aide à la rédaction de *data papers* proposées par des ingénieurs notamment à l'Université de Lorraine où la science ouverte tient une place importante (Bracco, 2022). L'écriture du *data paper* constitue une opportunité pour les personnels d'accompagnement à la recherche de valoriser leur rôle. Le travail sur les données est difficilement appréciable extérieurement, la micro-attribution permet à des utilisateurs extérieurs au projet de connaître les auteurs de telle ou telle expérience, qui sont souvent les « petites mains » de la recherche (Waquet, 2022). Ouvrir la science et montrer d'une manière transparente qui a fait quoi à chaque étape d'un projet permet de s'inscrire réellement dans une démarche de reproductibilité : l'auteur de l'expérience est *a fortiori* le plus à même de répondre aux questions futures sur l'expérience.

41. Toutefois, les problématiques liées à cette co-construction pourraient être résolues par la possibilité de générer automatiquement un *data paper*. L'écriture d'un *data paper* est donc très rarement une œuvre solitaire, mais un processus où le collectif apparaît à plusieurs étapes pour la gestion des données dont il est question, qui n'est cependant pas toujours mentionné dans les auteurs de l'article. À travers l'adoption du *data paper* par la communauté scientifique, il s'agit d'accepter que la recherche ne soit pas (plus) un chercheur qui soumet un article à un journal, mais bien un ensemble d'acteurs au sein d'une structure étendue construite autour des données de recherche d'un projet, de leur production et curation jusqu'à l'archivage.

42. Un dernier point doit être souligné lorsqu'on évoque la rédaction d'un *data paper*, celui des mises à jour. En effet, dans l'idéal, le *data paper* devrait tenir compte du versionnage des données et de la modification de leur description (Kunze *et al.*, 2011), et devrait, contrairement à l'article traditionnel, pouvoir être mis à jour, avec dans ce cas différentes versions du document afin d'assurer une reproductibilité à un moment donné.

L'évaluation

43. Dans la mesure où les *data papers* sont publiés dans des revues scientifiques, ils doivent faire l'objet d'une sélection par les pairs, en d'autres termes, ils doivent être évalués. Cependant, la réalité est plus nuancée, pour plusieurs raisons.

44. Commençons par l'aspect qui fâche : comme le modèle dominant de ces revues est le libre accès avec APC, le risque de ce qu'on a appelé le *predatory publishing* – la publication très rapide d'une grande partie des manuscrits sans véritable évaluation – est réel, notamment (mais pas exclusivement) parmi les éditeurs 100 % *open access*. Ceci étant, il n'y a pas eu d'analyses des revues de données sous cet angle, et il n'est pas possible de dire s'il y a vraiment des revues prédatrices parmi elles ou pas.

45. Le deuxième point est lié à la nature même des *data papers* : avec une structure (souvent) rigide (des « cases à remplir ») et sans présentation ni discussion des résultats d'analyse, la nécessité et l'importance d'une évaluation approfondie des *data papers* ne sont pas les mêmes que pour des articles de recherche. Les enjeux ne sont pas du même ordre non plus. Publier des résultats faux ou erronés peut avoir des effets délétères ; mais quel serait l'impact d'un *data paper* de mauvaise qualité ?

46. Ajoutons un troisième aspect, par rapport au processus et à la fonction de ces publications : nous avons évoqué plus haut que les *data papers* sont davantage liés au processus de la gestion des données et qu'ils n'ont pas de valeur intrinsèque, puisque leur fonction principale est d'indiquer l'endroit où se trouvent les données, de préciser les conditions d'accès et d'usage et d'en faciliter la réutilisation par une description détaillée. Ils dépendent en partie au moins des plans de gestion et des métadonnées des données, de leur indexation dans un entrepôt. Il s'agit beaucoup moins d'une création originale que d'une reprise et d'un enrichissement d'informations produites dans le cadre du cycle de la vie des données. L'intérêt d'une évaluation est donc moindre, comparé à celle d'articles de recherche.

47. Dans l'analyse de revues de données, à l'exception d'un titre, tous les *data journals* ont mis en place une procédure de sélection sous forme de *peer review*, souvent par deux experts (Schöpfel *et al.*, 2019 ; cf. aussi Seo et Kim, 2020). *Scientific Data* précise qu'un des évaluateurs est un expert dans le domaine des données (*data standards expert*) tandis que l'autre est un expert du domaine scientifique. Trois constats :

1. Nous n'avons pas trouvé de revue de données avec *peer review* en « double aveugle », uniquement des revues en « simple aveugle » (les auteurs ne connaissent pas l'identité des évaluateurs). Par exemple, la revue *IUCrData* pratique une évaluation en simple aveugle par au moins deux experts et indique que les manuscrits non acceptés après deux tours d'évaluation ne seront pas publiés.
2. Une autre revue (*Chemical Data Collections*) applique un « examen rapide par les pairs » en se concentrant sur la valeur des données et leur réutilisation potentielle, mais n'explique pas qui procède à l'examen par les pairs et combien de temps cela prend.
3. Cinq revues mentionnent une forme d'*open peer review*, soit en option, soit pour tous les *data papers*. Quelques exemples : les auteurs de *F1000Research* connaissent (et peuvent suggérer) leur *reviewers* ; le *Biodiversity Data Journal* fait évaluer les manuscrits par la com-

munauté ; *Earth System Science Data* a mis en place un examen public interactif par les pairs. La revue *Data* indique que les évaluateurs peuvent choisir de signer leurs évaluations et que les auteurs peuvent choisir d'inclure les rapports des évaluateurs comme matériel supplémentaire.

48. Cette dernière revue est particulièrement intéressante : les avis des pairs et des rédacteurs, les réponses des auteurs, ainsi que les différentes versions du manuscrit sont rendus publics avec le *data paper*. Par ailleurs, comme la valeur des données n'est pas vraiment connue avant qu'elles ne soient largement utilisées et annotées, leur évaluation devrait intervenir *après* la publication et le partage des données, donc, par exemple, en même temps que les *data papers* (Kunze *et al.*, 2011).

49. On peut résumer ce constat en deux mots : il s'agit globalement d'une évaluation allégée qui exprime un souci de transparence. Apparemment, les *data papers* sont évalués, mais pas tout à fait de la même manière que les articles de recherche, ce qui ne semble pas surprenant, pour des raisons indiquées plus haut. Reste une question : évaluation et sélection, oui, mais pourquoi ? Pour assurer et garantir la qualité et la réputation de la revue, comme pour les articles « normaux » où un taux de sélection élevé est considéré comme indicateur de qualité de la revue ? Pour augmenter l'impact des données, en termes d'accès et de réutilisation ? Ou, troisième option, pour contribuer à la crédibilité (*trustworthiness*) des entrepôts dans lesquels se trouvent les données signalées ? Question ouverte.

La qualité

50. Quels sont les critères d'évaluation ? Les deux revues les plus importantes, *Data in Brief* et *Scientific Data* appliquent les critères suivants :

1. La conformité du format des données aux normes existantes (alignement avec les normes communautaires).
2. Une explication suffisante du protocole/des références pour générer les données (qualité technique des procédures).
3. L'exhaustivité de la description des données (qualité de la documentation).
4. Une explication adéquate de l'utilité des données.
5. La valeur de réutilisation des données.
6. La conformité de l'article au modèle (instructions aux auteurs).

51. Dans d'autres revues, comme le *Geoscience Data Journal*, l'évaluation porte sur la description des données, sur la qualité des métadonnées et sur les données elles-mêmes (accessibilité, facilité d'utilisation…). Il y a donc une double vérification, scientifique d'une part (méthodologie, intérêt, etc.), technique d'autre part (métadonnées, accès, etc.). Normalement, les deux aspects sont évalués par les mêmes personnes, à l'exception de la revue *Giga-Science* qui confie l'évaluation technique des données à un *data reviewer* qui vérifiera aussi si les données correspondent à la description (Walters, 2020 ; Mayernik *et al.*, 2015). D'autres critères sont plus classiques, comme la qualité des illustrations, la lisibilité et la clarté du style, la structuration, etc. Dans le domaine de la médecine, *Open*

Health Data invite les *peer reviewers* à évaluer si l'entrepôt de données permet la protection des données sensibles.

52. Quelques revues, à l'instar de *Scientific Data*, précisent que l'acceptation n'est pas basée sur l'impact ou la nouveauté des résultats, mais seulement sur l'intérêt scientifique des données. D'autres, en revanche, évoquent quand même ce critère de nouveauté (Kratz et Strasser, 2015).

53. L'évaluation d'un *data paper* ne porte donc pas uniquement sur la qualité du manuscrit, mais va plus loin, dans la mesure où une partie des critères concerne directement la qualité des données et (moins souvent) le choix de l'entrepôt et la garantie d'une protection et d'un archivage pérenne des données.

L'impact

54. Que savons-nous de l'impact des *data papers* ? Quelques études montrent qu'environ deux tiers des *data papers* sont cités par d'autres articles, mais que le nombre de citations est généralement plus faible que pour les articles de recherche, avec un effet de longue traîne (Kotti *et al.*, 2020 ; Adamczak, 2021). Le fait que presque tous les *data papers* soient publiés en libre accès doit contribuer à cet impact.

55. Le taux d'articles cités et le nombre moyen de citations par article varient d'une revue à l'autre. Aussi, les revues de données les plus importantes ont déjà un facteur d'impact ou sont en train de l'acquérir (Adamczak, 2021).

56. D'autres études sont plus réservées, évoquant un impact plutôt faible de ce nouveau type de publication (Thelwall, 2020) et un intérêt limité pour les auteurs (Huang et Jeng, 2022). Mais au fond, la vraie question n'est pas le nombre de citations des *data papers* ; dans la mesure où leur fonction principale est le signalement des données afin de favoriser leur réutilisation, la question de l'impact des *data papers* se pose différemment : dans quelle mesure les *data papers* contribuent-ils à la consultation des données et à leur réutilisation ? En d'autres termes, y a-t-il un rapport entre la réutilisation des données dans les entrepôts et la publication des *data papers* sur les plateformes de revues ?

57. Et sur ce point précis, il n'y a, à ce jour, que peu d'évidence empirique (Jiao et Darch, 2020). L'analyse de la revue *Data in Brief* aboutit au constat que, dans l'ensemble, la revue semble apporter une contribution positive à la science en permettant l'accès aux données, mais que ses *data papers* débouchent rarement sur une réutilisation des données (Thelwall, 2020). De l'autre côté, Wang *et al.* (2021) observent que la réutilisation des données est partiellement conditionnée par l'existence et la qualité d'un *data paper*, de même qu'une documentation insuffisante et le manque d'accessibilité font obstacle à leur réutilisation.

58. Une petite étude des données du réseau international GreyNet a établi ce rapport : dans l'entrepôt de données en SHS *EASY* (Pays-Bas), les données avec un *data paper* ont été davantage consultées et téléchargées que les données sans *data paper*, à un niveau significatif (Farace et Schöpfel, 2020). Toujours en SHS, l'analyse de deux revues de données (*Journal of Open Humanities Data*, *Research Data Journal for the Humanities and Social Sciences*) confirme le constat : apparemment, les *data papers* ont un impact positif à la fois sur la citation des articles de recherche qui leur sont associés et sur la réutilisation des données (McGillivray *et al.*, 2022). Trois remarques pour terminer :

1. Même s'il existe encore peu d'études sur l'impact des données de recherche en SHS, leur taux de citation ne semble pas très élevé (Robinson-Garcia *et al.*, 2016).
2. Les études qui ont partagé les données dans un entrepôt sont davantage citées que les autres (Piwowar et Vision, 2013).
3. Dernière remarque : comment mesurer l'impact des données dans un entrepôt (Khan *et al.*, 2021) ? Est-ce que le fait de télécharger et/ou de citer un jeu de données veut dire « réutilisation » ?

59. En fin de compte, nous ne sommes qu'au tout début d'un écosystème d'articles scientifiques, de données de recherche et de *data papers* dont nous commençons seulement à comprendre les éléments, les relations et les dynamiques.

Révéler les formes et logiques de citation des *data papers* en archéologie : le cas du *Journal of Open Archaeology Data*

Violaine Rebouillat

1. Deux questions reviennent quand on parle du partage des données : celle des conditions de leur réutilisation et celle de la reconnaissance du travail investi. Depuis le début des années 2010, le *data paper* est proposé comme l'un des outils capables de répondre à ces deux enjeux, de plus en plus présents dans le champ de la communication scientifique sous l'effet des politiques de science ouverte. Premièrement, le *data paper* a pour fonction de faciliter la réutilisation des données, dont il est complémentaire, grâce à une description fine du contexte de production (Callaghan *et al.*, 2012 ; Candela *et al.*, 2015 ; Schöpfel *et al.*, 2019). Deuxièmement, il s'appuie sur le modèle traditionnel de l'article scientifique, dont il mobilise les mécanismes de reconnaissance (évaluation par les pairs, publication, paternité, citation) (Seo et Kim, 2020 ; Walters, 2020).

2. Ce type de publication, reposant sur une description relativement libre du jeu de données, pourrait sembler particulièrement adapté aux Sciences humaines et sociales (SHS), dont les données se caractérisent souvent par des contextes de production complexes. Or, force est de constater qu'à l'heure actuelle, les *data papers* et *data journals* en SHS ne connaissent pas le même essor que dans les Sciences, Techniques et Médecine (STM) (Walters, 2020).

3. Une discipline semble pourtant se distinguer : il s'agit de l'archéologie, qui fait figure de précurseur en matière d'offre éditoriale de *data journals*. Dans l'étude de Schöpfel *et al.* (2019), deux des quatre *data journals* identifiés en SHS étaient ainsi spécialisés en archéologie.

4. Cette avance sur les autres domaines nous permet d'étudier l'archéologie comme discipline exploratoire de l'émergence des *data papers* en SHS. Le présent chapitre se concentre sur l'utilisation des *data papers*, dont il observe l'impact au travers de leur citation dans la littérature scientifique. Notre but est de comprendre les logiques suivantes : comment se caractérisent les pratiques de citation de *data papers* en archéologie ? Comment ces pratiques sont-elles justifiées par leurs auteurs ? Quelles sont les motivations de ces derniers ?

5. Ces questions ont peu été investies par la littérature scientifique. Il n'existe à ce jour aucune publication connue sur les *data papers* en archéologie. En revanche, plusieurs travaux se sont penchés sur les enjeux du partage des données (Borgman, 2015 ; Huvila, 2016 ; Kansa et Kansa, 2018), l'archéologie étant, par la nature même de ses objets et méthodes, une discipline sensible à la valeur de préservation (Frank, Yakel, et Faniel, 2015). Ces

travaux nous permettent donc de prendre en compte la spécificité des données archéologiques.

6. Quant aux pratiques de citation des *data papers*, elles ont fait l'objet de quelques études en STM, révélant des spécificités disciplinaires (Jiao et Darch, 2020) ainsi qu'une part importante d'auto-citations (Kotti et Spinellis, 2019). Notre recherche propose d'enrichir ces travaux en déployant une approche centrée sur le chercheur, qui permette de mettre au jour les logiques de citation des *data papers*. Elle étudiera les *data papers* en tant qu'articles rédigés par des auteurs, évalués par des pairs et publiables dans des revues scientifiques, dont le contenu principal est une description de jeux de données de recherche publiés, ainsi que de leur contexte de production et d'acquisition, selon la définition proposée par Schöpfel *et al.* (2019).

7. Notre étude porte sur un *data journal* de langue anglaise en archéologie, que nous prenons comme cas d'étude : le *Journal of Open Archaeology Data* (*JOAD*). Créé en 2012 et hébergé par l'éditeur britannique Ubiquity Press, le *JOAD* s'appuie sur le modèle éditorial de l'article scientifique, à savoir une évaluation par les pairs et une diffusion en Gold Open Access[1]. Il couvre tous types de données du domaine de l'archéologie, à la condition que celles-ci aient été au préalable déposées en libre accès dans un entrepôt de données. Plusieurs entrepôts sont recommandés par la revue[2], dont des entrepôts du domaine de l'archéologie et des entrepôts multidisciplinaires, sélectionnés pour la pérennité de leur modèle. À la date de début de notre étude (le 26 mai 2021), 41 *data papers* avaient été publiés dans le JOAD.

8. L'approche méthodologique repose sur une enquête menée auprès des auteurs ayant cité un des *data papers* du *JOAD*. Elle vise à comprendre leurs motivations et à identifier les formes de citation mobilisées (objet de la citation, motifs, nature du lien à l'auteur).

9. Ce faisant, l'originalité de cette recherche consiste à proposer un regard issu des Sciences de l'information et de la communication (SIC) sur les Humanités numériques, ce que Boukacem-Zeghmouri et Paquienséguy (2021) ont appelé un « devoir de critique ». En mobilisant les outils des SIC, nous entendons apporter un éclairage sur les logiques des chercheurs en tant que lecteurs et auteurs, aux prises avec les mutations de l'édition et des données scientifiques numériques en SHS.

1. Le modèle économique du *JOAD* repose sur une diffusion des articles en libre d'accès. Les coûts liés à la publication (processus éditoriaux, hébergement web, indexation, marketing, archivage, enregistrement DOI…) sont couverts par les auteurs, auxquels il est demandé des frais de publication (APC) de 350 dollars. Dans le cas où l'auteur ne disposerait pas des fonds nécessaires pour payer les frais de publication, une réduction ou une exonération peut toutefois être envisagée.

2. Des entrepôts institutionnels (JOAD Dataverse, Archaeology Data Service, Figshare, Open Context, tDAR, Zenodo) ; des entrepôts nationaux (Arachne, ARCHE, DANS, Mappa, SND) ; des entrepôts institutionnels (UCL Discovery). Source: https://openarchaeologydata.metajnl.com/about/#repo.

Méthodologie

10. La méthodologie déployée repose sur la réalisation d'une enquête par questionnaire, diffusée aux auteurs ayant cité un des *data papers* du *JOAD* (soit 43 auteurs correspondants pour un total de 50 articles citants). L'objectif est de comprendre les raisons et motivations de ces auteurs à citer un *data paper* et d'analyser s'il s'agit d'une pratique isolée ou bien, au contraire, d'habitudes plus ancrées.

11. La construction de l'enquête repose sur deux grandes étapes successives.

Identification de l'échantillon

12. La première étape a consisté à identifier l'échantillon ciblé pour l'enquête. Il s'est agi, dans un premier temps, de recenser les publications dans lesquelles étaient cités un ou plusieurs *data papers* du *JOAD*. La liste des publications en question a été dressée à partir des données fournies sur le site web du *data journal*, où sont référencés le nombre de citations reçues par chaque *data paper* ainsi que les références bibliographiques associées. Ces données ont été recoupées avec celles du Web of Science, du moins pour les 31 *data papers* du *JOAD* qui y sont référencés. Un total de 50 publications a ainsi pu être répertorié.

13. Pour chacune d'entre elles, nous avons ensuite récupéré le courriel de l'auteur correspondant sur le site de l'éditeur. Au total, nous avons répertorié 43 auteurs différents, soit un nombre plus petit que celui des publications citantes, certains auteurs ayant cité plusieurs *data papers* dans une publication, d'autres ayant cité un *data paper* dans plusieurs publications.

Élaboration et diffusion du questionnaire

14. Un questionnaire en ligne, destiné à sonder cette population, a alors été conçu. Composé de 38 questions, il était structuré en trois parties. La première avait pour objectif d'établir le profil des répondants, avec des questions portant sur le statut, l'âge, la discipline, l'affiliation et le pays d'exercice. La deuxième partie interrogeait les auteurs sur leur connaissance des *data papers* et leurs motivations à citer un des articles du *JOAD*. La troisième et dernière partie visait à connaître leurs pratiques de citation de *data papers* au-delà de cette expérience et à recueillir leur opinion sur leur pertinence dans un domaine comme l'archéologie. Au niveau de la question 7, le questionnaire se divisait en deux voies, l'une réservée aux auteurs ayant cité leur propre *data paper*, l'autre destinée aux auteurs ayant cité le *data paper* d'un autre chercheur, avec pour chacune des questions adaptées (respectivement 11 et 20 questions).

15. L'enquête a été créée avec l'outil SurveyMonkey. Elle a été diffusée aux 43 auteurs correspondants entre le 15 juillet et le 17 août 2021. Quatre courriels se sont avérés invalides. Au terme de la phase de diffusion,

l'enquête avait reçu 22 réponses, dont 3 incomplètes que nous avons dû exclure de l'analyse. Les résultats présentés ici se concentrent donc sur un ensemble de 19 réponses exploitables. Étant donné la taille restreinte de la population ciblée dans l'enquête, l'objectif n'est pas de proposer une analyse statistique des données, mais plutôt d'étudier sous un angle individuel et qualitatif les réponses des répondants.

16. Il faut préciser que la construction du questionnaire repose sur la réalisation en amont d'une analyse bibliométrique des données du *JOAD*. Celle-ci visait à étudier les citations reçues par les *data papers* de la revue, afin d'estimer quantitativement leur importance et établir un premier profil des auteurs et types de publication citants.

Un intérêt pour le partage des données archéologiques

Profil des répondants

17. Le panel est composé de répondants aux statuts très variés, allant du doctorat à l'éméritat, avec une majorité de chercheurs âgés de 31 à 45 ans (figure 1). Ils viennent des pays occidentaux (tableau 1) et sont pour la plupart (16) affiliés à des universités (tableau 2).

18. Ils appartiennent tous au domaine des Sciences humaines et sociales, à l'exception d'un répondant en biologie. Cinq d'entre eux ont précisé qu'ils travaillaient plus spécifiquement dans le champ de l'archéologie.

19. Les répondants se divisent en deux profils : ceux qui ont cité leur propre *data paper* (11 répondants) et ceux qui ont cité le *data paper* d'autres chercheurs (8 répondants).

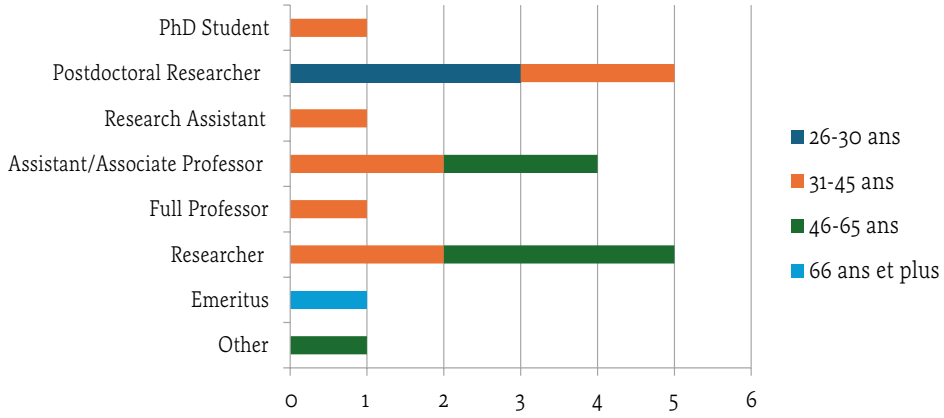


Figure 1. Statut des répondants

	Nombre de répondants
Royaume-Uni	4
France	2
Allemagne	2
Espagne	2
Danemark	2
États-Unis	2
Nouvelle-Zélande	2
Norvège	1
Italie	1
Grèce	1

Tableau 1. Pays d'exercice

	Nombre de répondants
Aarhus University	1
Arizona State University	1
British Museum	1
CNRS	2
Ludwig Maximilian University of Munich	1
Museum of Cultural History	1
Sapienza University of Rome	1
University of Auckland	2
University of Cambridge	3
University of Cantabria	1
University of Chicago	1
University of Copenhagen	1
University of La Laguna	1
University of Patras	1
University of Tuebingen	1

Tableau 2. Établissement d'affiliation

L'importance accordée aux *data papers*

20. L'une des questions visait à interroger les répondants sur la pertinence des *data papers* en archéologie. Les avis recueillis sont globalement positifs (17 répondants), 2 sont plus mitigés. Pour la majorité des répondants, le *data paper* constitue un support de communication essentiel au processus de publication des données. Leurs réponses se basent sur trois arguments que nous reportons ici.

Premier argument

21. Le premier argument met en évidence l'importance des *data papers* pour contextualiser le processus de collecte et de traitement des données archéologiques.

22. « En archéologie, les théories reposent de plus en plus sur des jeux de données numériques qui peuvent être compilés à l'aide de différentes méthodologies impliquant de nombreuses transformations ultérieures. La réutilisation des données et la reproductibilité scientifique nécessitent la mise à disposition de ces jeux de données numériques compilés et leur description afin de comprendre le processus de compilation qui a été appliqué. »[3]

23. Le *data paper* permet d'expliciter la méthodologie déployée et en particulier de retracer le processus décisionnel ayant conduit au choix de collecter telles ou telles données, comme le précise un des répondants :

24. « [Les *data papers*] sont des éléments essentiels, car les données archéologiques sont rarement normalisées et de nombreuses décisions sont prises quant à quelles données collecter. Le *data paper* est la seule source qui rend ces décisions transparentes. »[4]

3. « In archaeology arguments are increasingly based on digital datasets that can be compiled using different methodologies involving many subsequent transformations. Both data reuse and scientific reproducibility require the availability of compiled digital datasets and their description to understand the dataset compilation process. »

4. « [Data papers] are critical components, as archaeological data is rarely standardised with many decisions that are made about what data to collect. The data paper is the only source that makes those decisions transparent. »

Deuxième argument

25. Les *data papers* ont également pour avantage de prolonger les articles de recherche en permettant de décrire plus finement les données et en donnant ainsi les informations nécessaires à la vérification et à la reproductibilité des résultats. Le répondant ci-dessous explique que les *data papers* sont d'autant plus utiles lorsque le jeu de données est complexe et qu'il ne peut être présenté de manière suffisamment détaillée sans dépasser le nombre maximum de caractères dévolus à l'article :

26. « Les articles de revue devenant le principal moyen de publication (et non les livres), il est logique que des *data papers* spécialisés soient désormais publiés également. Parfois, les *data papers* sont la seule source d'examen empirique au sein de la communauté scientifique, car certains types de données sont trop complexes pour être diffusés dans les canaux de publication habituels. Si nous ne pouvons pas avoir accès aux données, comment pouvons-nous vraiment faire confiance de manière scientifique à une conclusion dépendante des données ? »[5]

Troisième argument

27. Le troisième et dernier argument porte sur le partage et la réutilisation des données, auxquels les *data papers* contribuent. Pour cinq des répondants, les *data papers* constituent un moyen de favoriser et d'accroître le partage des données en archéologie :

28. « J'ai trouvé que le *data paper* était l'une des introductions les plus pertinentes dans le domaine de l'archéologie. Notre domaine est malheureusement rempli de publications où le partage des données est rare, voire complètement absent, surtout en archéologie du Proche-Orient. Ce n'est pas toujours la faute des chercheurs eux-mêmes, mais l'introduction des *data papers* et leur importance évidente permettront, je l'espère, de lancer une nouvelle tendance au partage des données, ce qui est fondamental dans tout domaine de recherche. »[6]

29. « Je trouve le *data paper* extrêmement utile, car, dans mon domaine (l'archéologie), l'accès aux données sources est encore très limité et le partage des données n'a fait que récemment son entrée dans les publications. Bien que je considère qu'un *data paper* soit différent d'un article de recherche, je pense qu'il s'agit d'un aspect complémentaire de tout article de recherche. »[7]

5. « As journal publications are becoming the primary way to publish (and not books), it makes sense that specialized *data papers* are now being published also. Sometimes, *data papers* are the only source for empirical scrutiny within the scientific community, because some types of data are too complex for dissemination through regular publication channels. If we cannot see the data, how can we truly trust a data dependent conclusion in research? »

6. « I found data paper to be one of the most relevant introductions in the field of archaeology. Our field is unfortunately full of publications where the sharing of data is minimal if not completely absent, especially in Near Eastern Archaeology. This is not always the fault of the researchers themselves, but the introduction of *data papers* and their obvious importance will hopefully start a new trend of data sharing, which is fundamental in any research field. »

7. « I find the data paper extremely useful as in my field (Archaeology), access to source data is still very limited and data sharing is only recently made its way into the publi-

30. Cet argument s'inscrit, comme le précise l'un des répondants, « à l'heure de la sensibilisation à la science ouverte »[8]. Le contexte d'ouverture contribue donc probablement à l'orientation de certaines réponses.

31. Au-delà de ce contexte, un des répondants a insisté sur l'importance particulière de l'accès aux données en archéologie, la réanalyse et la compilation de données antérieures étant au cœur des pratiques de la discipline : « [les *data papers* sont] très pertinents, car les améliorations dans cette discipline reposent en grande partie sur l'accès aux données primaires et sur leur réinterprétation »[9]. La nature du travail exercé sur les données déterminerait donc l'utilité plus ou moins grande du *data paper*.

32. Ces arguments nous amènent à faire l'hypothèse que la population interrogée rassemble un profil de chercheurs bien spécifique, engagés dans l'ouverture des résultats de la science ou du moins sensibles à l'importance de rendre plus accessibles les données de la recherche archéologique. Les répondants semblent en effet partager un même intérêt pour l'accès aux données archéologiques, intérêt qui les a probablement conduits à publier et/ou citer ce type de publication. Les termes de « partage des données », d'« accès aux données », de « reproductibilité des données », voire d'« *open science* » sont souvent revenus dans leurs réponses. L'un des articles citant un *data paper* portait d'ailleurs sur la question de la réutilisation des données.

33. L'enquête révèle néanmoins un autre aspect des *data papers*, à savoir leur introduction récente dans le domaine de l'archéologie. Le terme « *data paper* » apparaît en effet comme un terme nouveau. Son utilisation n'est pas systématique et sa définition pas toujours maîtrisée, y compris dans notre panel d'auteurs ayant déjà publié et/ou cité un *data paper*. Si le terme est utilisé par la plupart d'entre eux (15 répondants), il reste encore méconnu par d'autres (4 répondants), qui ne font pas la différence de manière évidente entre un article de recherche classique et un *data paper* (ils utilisent le terme générique d'« article » pour désigner leur *data paper*). L'un d'eux semble confondre *data paper* et jeu de données (« de mémoire, le terme actuel est *big data* »[10]). Ce type de confusion se retrouve chez les deux catégories de répondants, aussi bien chez ceux qui ont publié un *data paper* que chez ceux qui ont uniquement cité un *data paper*. Elle est le signe qu'il s'agit d'un type de publication encore émergent dans la communauté des archéologues.

cations. While I do consider a data paper different from a research paper, I believe it is a complementary aspect of any research paper. »

8. « at the time of heightened awareness about open science »

9. « [Data papers are] very relevant since improvements in that discipline are largely based on the access to – and reinterpretation of – primary data. »

10. « From memory the actual term is big data. »

La citation de *data papers* en archéologie, une pratique émergente ?

34. Ce statut émergent du *data paper* se retrouve dans les pratiques de citation. L'analyse bibliométrique réalisée en amont de l'enquête montre que les articles du *JOAD* sont encore peu, voire pas cités. Sur les 41 *data papers* publiés dans la revue à la date du 26 mai 2021, 17 avaient déjà été cités au moins une fois. Parmi eux, le nombre de citations varie de 1 à 13, avec une moyenne de 3 citations par *data paper* (tableau 3). L'analyse ne montre pas de corrélation entre la date de publication du *data paper* et le nombre de citations reçues.

35. Les publications dans lesquelles ces *data papers* sont cités sont unilatéralement issues du domaine de l'archéologie. Elles rassemblent des *data papers* (4) et des articles de revue (46).

36. Les citations reçues proviennent à la fois d'auto-citations (20) et de citations par les pairs (30). Les auto-citations représentent en moyenne 38,5 % du total des citations par *data paper*, une part relativement élevée comparée à celle des articles de revues classiques. Une étude de Hutson (2006), issue de l'analyse de quatre revues en archéologie, donnait ainsi un nombre moyen d'auto-citations de 8,4 %. Cette différence pourrait s'expliquer par le caractère émergent des *data papers*, faisant d'eux des objets encore peu connus et donc peu cités, si ce n'est par leurs propres auteurs.

37. Les résultats du questionnaire étayent et nuancent ces premiers constats. Du côté des chercheurs, en effet, la citation d'un *data paper* ne semble pas s'arrêter à une expérience unique. Sur 10 des répondants à l'enquête, 7 déclarent citer des *data papers* de temps en temps et 3 rarement. De même, les 11 auteurs ayant cité leur propre *data paper* ne se limitent pas à des pratiques d'auto-citation : 8 d'entre eux ont en effet déclaré avoir déjà cité des *data papers* d'autres chercheurs. Ces pratiques de citation de *data papers* sont toutefois relativement récentes : les plus anciennes citations du *JOAD* remontent à 2014 et, pour la majorité des répondants à l'enquête (13), elle correspondait à leur première expérience de citation d'un *data paper*.

38. La fréquence de citation de *data papers* dépend aussi du sujet des *data papers* publiés. Comme l'écrit l'un des répondants, « lorsque les *data papers* sont pertinents pour mon sujet de recherche ou lorsque je réutilise les données, je cite les *data papers* (s'ils existent et si j'en ai connaissance) »[11]. À l'heure de l'émergence des *data papers*, il n'apparaît pas évident pour un chercheur en archéologie de trouver des *data papers* proches de son objet ou de son sujet de recherche. Un autre répondant justifiait qu'il citait rarement des *data papers* pour cette même raison : « il y en a peu qui étaient pertinents pour mes recherches »[12].

11. « When *data papers* are relevant to the topic or when I reuse the data, I cite *data papers* (if they exist and I know about them). »

12. « There were not many that were relevant to my work. »

Data paper Du JOAD	DOI	Nombre de citations (26/05/2021)	Dont nombre d'auto-citations (26/05/2021)
Data paper 1	10.5334/4f293686e4d62	0	0
Data paper 2	10.5334/4f33a7b040dd1	0	0
Data paper 3	10.5334/4f7b093ed0a77	0	0
Data paper 4	10.5334/4f7db511ae16c	0	0
Data paper 5	10.5334/4f8eb4078284b	0	0
Data paper 6	10.5334/4f8d6ed49bd54	0	0
Data paper 7	10.5334/4f913ca0cbb89	0	0
Data paper 8	10.5334/joad.aa	0	0
Data paper 9	10.5334/joad.ab	0	0
Data paper 10	10.5334/joad.af	0	0
Data paper 11	10.5334/joad.ag	0	0
Data paper 12	10.5334/joad.ah	0	0
Data paper 13	10.5334/joad.44	0	0
Data paper 14	10.5334/joad.51	0	0
Data paper 15	10.5334/joad.53	0	0
Data paper 16	10.5334/joad.57	0	0
Data paper 17	10.5334/joad.56	0	0
Data paper 18	10.5334/joad.60	0	0
Data paper 19	10.5334/joad.61	0	0
Data paper 20	10.5334/joad.62	0	0
Data paper 21	10.5334/joad.67	0	0
Data paper 22	10.5334/joad.71	0	0
Data paper 23	10.5334/joad.68	0	0
Data paper 24	10.5334/joad.72	0	0
Data paper 25	10.5334/joad.ac	1	0
Data paper 26	10.5334/joad.ae	1	0
Data paper 27	10.5334/joad.aj	1	1
Data paper 28	10.5334/joad.45	1	1
Data paper 29	10.5334/joad.59	1	0
Data paper 30	10.5334/joad.65	1	0
Data paper 31	10.5334/joad.ad	2	1
Data paper 32	10.5334/joad.49	2	0
Data paper 33	10.5334/joad.63	2	1
Data paper 34	10.5334/4f60dd6baa298	3	1
Data paper 35	10.5334/joad.ai	3	2
Data paper 36	10.5334/joad.41	3	1
Data paper 37	10.5334/joad.52	3	3
Data paper 38	10.5334/4f3bcb3f7f21d	4	0
Data paper 39	10.5334/joad.42	4	1
Data paper 40	10.5334/joad.43	5	3
Data paper 41	10.5334/joad.40	13	5

Tableau 3. Citations reçues par les *data papers* du JOAD

Motivations à citer un *data paper* du *JOAD*

39. Si les pratiques de citation s'avèrent encore peu développées, il est intéressant d'explorer les raisons et les motivations qui ont conduit quelques rares auteurs à citer un *data paper* du *JOAD*. Tel était l'objet de la deuxième partie du questionnaire. Nous distinguons ici

les retours des auteurs ayant cité leur propre *data paper* des autres auteurs.

Réponses des auteurs ayant cité leur propre *data paper*

40. Les 11 auteurs ayant cité leur propre *data paper* ont invoqué des motifs de types divers. Trois d'entre eux ont été précis sur l'utilisation du *data paper* dans leur publication : deux ont expliqué qu'ils avaient fait une nouvelle analyse des données décrites dans le *data paper*, tandis que le troisième a déclaré s'en être servi comme exemple dans un article sur la réutilisation des données.

41. Les autres répondants ont plutôt axé leur réponse sur les avantages apportés à leur travail de rédaction d'une publication. Pour deux d'entre eux, citer le *data paper* leur a permis de faire appel à des données sans avoir à les décrire de zéro. Un autre a indiqué qu'il était plus facile pour lui de citer le *data paper* que chacun des articles de recherche traitant des données. Le *data paper* est ainsi perçu par ces auteurs comme une entité citable facilitant la rédaction de leurs publications.

42. Une autre question dans l'enquête permettait de connaître plus spécifiquement les bénéfices recherchés par les auteurs en citant leur *data paper*. Parmi les réponses proposées, celles qui sont revenues le plus souvent sont : « valoriser les efforts investis dans la collecte et le traitement des données » et « encourager le partage et la réutilisation des données décrites dans le *data paper* », devant le fait d'accroître la visibilité du *data paper* ou de promouvoir ce type de publication (figure 2). La première réponse (« valoriser les efforts investis dans la collecte et le traitement des données ») met l'accent sur le besoin de reconnaissance du travail de production des données, un travail qui, dans le domaine de l'archéologie, peut prendre plusieurs années et mobiliser plusieurs personnes (chercheurs, ingénieurs, doctorants…). Cette spécificité de la collecte des données, se caractérisant par un travail artisanal de longue haleine, en comparaison à d'autres approches où la collecte de données peut être automatisée, explique en partie pourquoi les données sont peu partagées et accessibles (Borgman, 2012). Le *data paper* permettrait alors de valoriser ce pan souvent invisibilisé du travail de recherche. La seconde réponse (« encourager le partage et la réutilisation des données décrites dans le *data paper* ») a davantage trait au profil « militant » des répondants, qui sont convaincus de l'intérêt de partager les données et qui souhaitent donc voir les données de leur *data paper* réutilisées (voir plus haut, partie « Un intérêt pour le partage des données archéologiques »).

Réponses des auteurs ayant cité le *data paper* d'autres chercheurs

43. Quant aux 8 auteurs ayant cité le *data paper* d'autres chercheurs, quatre questions permettaient de comprendre le contexte de citation du *data paper*. La

première question visait à connaître les éléments du *data paper* qui les avaient intéressés pour leur publication (les données, la méthode d'acquisition des données, le terrain d'où provenaient les données et/ou la façon dont les données avaient été compilées). La deuxième question portait sur les raisons ayant amené ces auteurs à citer le *data paper* et comportait trois propositions de réponses : le *data paper* fait partie intégrante de ma revue de littérature ; mon intention était de comparer mes données avec celles décrites dans le *data paper* ; j'ai utilisé les données du *data paper* dans mes recherches. Les deux dernières questions visaient à demander s'ils avaient eu besoin d'accéder au jeu de données décrit dans le *data paper* et s'ils l'avaient cité dans leur publication. Nous avons combiné les réponses obtenues à ces quatre questions, afin de mettre en lumière les différents contextes d'utilisation des *data papers* (tableau 4). Nous avons pu constater des cas de figure très variés, que nous avons souhaité illustrer d'exemples concrets, afin de comprendre de manière plus fine l'utilisation qui avait été faite des données et du *data paper*.

44. En complément, nous avons donc choisi de consulter les articles des cinq auteurs qui avaient laissé leur adresse mail dans l'enquête, afin de réaliser une analyse de contenu. Celle-ci nous a permis de comprendre le contexte de citation du *data paper* et de comparer ces résultats avec les réponses fournies par le chercheur. Parmi ces cinq auteurs, nous avons ainsi pu distinguer quatre motifs de citation : alimenter l'état de l'art ; compiler les données avec d'autres données ; comparer les données avec d'autres données ; analyser les données.

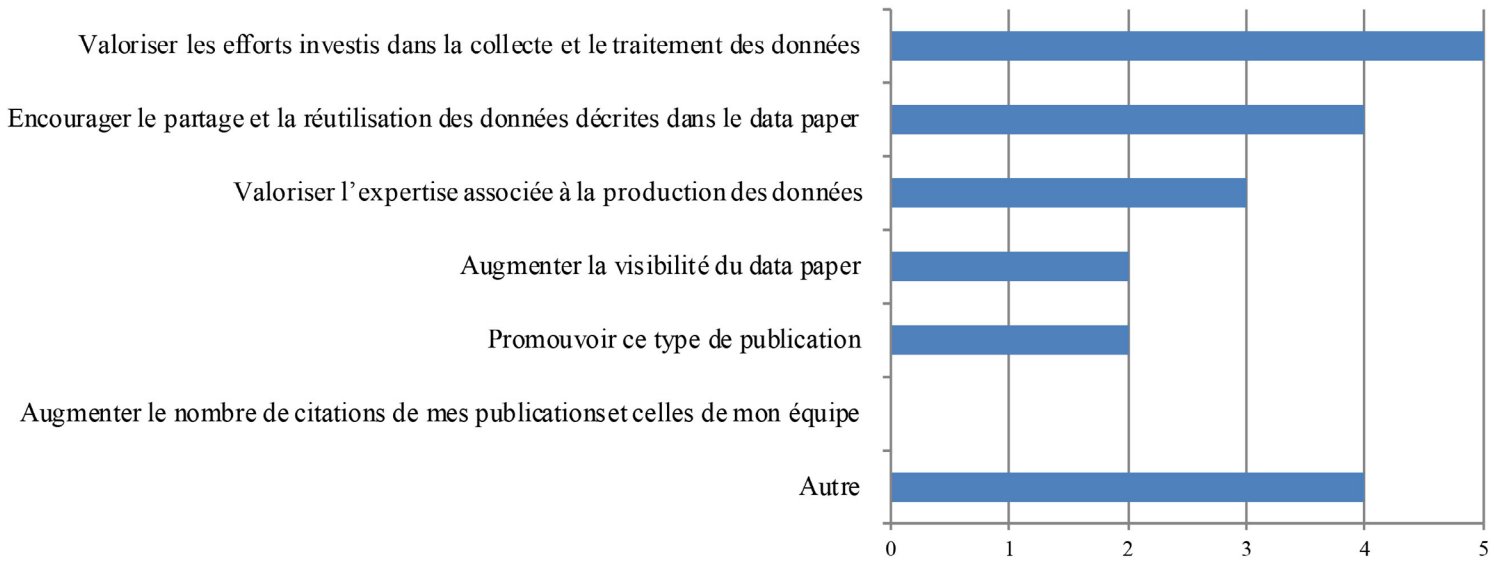


Figure 2. Bénéfices recherchés par la citation du *data paper*

	Quels éléments vous intéressaient dans ce *data paper* pour votre publication ?	Quel(s) motif(s) vous a/ont conduit à citer ce *data paper* ?	Avez-vous eu besoin d'accéder au jeu de données associé au *data paper* ?	Si oui, avez-vous également cité le jeu de données dans votre publication ?
Répondant 1	*Les données *La méthode d'acquisition des données *Le terrain dont les données sont issues *Le mode de compilation des données	*Le *data paper* faisait partie intégrante de mon état de la littérature. *Je souhaitais comparer mes données à celles décrites dans le *data paper*. *J'ai utilisé les données décrites dans le *data paper* dans mes recherches.	Oui	Oui
Répondant 7			Oui	Oui
Répondant 15			Oui	Non
Répondant 16		*Le *data paper* faisait partie intégrante de mon état de la littérature. *Je souhaitais comparer mes données à celles décrites dans le *data paper*.	Oui	Oui
Répondant 12	*Les données *La méthode d'acquisition des données	*J'ai utilisé les données décrites dans le *data paper* dans mes recherches.	Oui	Non
Répondant 5	*La méthode d'acquisition des données	*Le *data paper* faisait partie intégrante de mon état de la littérature.	Oui	Oui
Répondant 4		*Je souhaitais comparer mes données à celles décrites dans le *data paper*.	Non	/
Répondant 19	*Le terrain dont les données sont issues *Le mode de compilation des données		Non	/

Tableau 4. Contexte de citation du *data paper*
(compilation des questions 27 à 30)

Alimenter l'état de l'art

45. Dans le cas du répondant 4, le *data paper* a été cité dans l'état de la littérature de sa publication. L'élément du *data paper* qui l'intéressait était la méthode de collecte des données, qu'il a souhaité comparer avec d'autres techniques d'investigation. Cet élément nous permet de préciser sa réponse au questionnaire concernant la raison pour laquelle il a cité le *data paper* : l'objectif n'était pas de comparer les données décrites dans le *data paper* avec d'autres données, comme il le dit, mais de comparer plutôt la méthode d'investigation déployée. Le *data paper* lui a donc été utile à lui seul, il n'a pas eu besoin de consulter le jeu de données.

46. Nous avons retrouvé ce même type d'usage dans un article du répondant 1, où le *data paper* était cité dans l'état de l'art.

Compiler les données avec d'autres données

47. Le répondant 12 a analysé dans sa publication un jeu de données qu'il a constitué à partir de la compilation de trois bases de données préexistantes. L'une de ces trois bases est décrite dans un *data paper* du *JOAD*, que l'auteur cite dans la partie « Materials and Methods » de son article. Le *data paper* lui a permis de comprendre comment étaient structurées et renseignées les données dans la base de données et ainsi de prendre en main cette dernière. Le jeu de données, quant à lui, n'a pas été cité, même s'il est évident qu'il a été essentiel à ses recherches.

48. Le répondant 5 a quant à lui publié un *data paper*, dans lequel il cite un autre *data paper*. Tout comme le répondant 12, il a compilé plusieurs jeux de données existants en un seul. Le *data paper*, cité dans la partie « Methods », lui a permis de faire référence au jeu de données sous-jacent, là encore sans le citer directement.

49. Dans cette catégorie, on retrouve également le répondant 1 pour une autre de ses publications, dans laquelle il cite toujours le même *data paper*. Dans cette publication, les données du *data paper* ont été combinées avec d'autres jeux de données. Elles sont au cœur de l'article, avec un total de sept citations du *data paper* réparties dans les différentes parties de l'article (de l'introduction à la conclusion en passant par la méthodologie et les résultats). Là encore, seul le *data paper* est cité, alors qu'il y a bien eu un usage des données elles-mêmes.

Comparer les données avec d'autres

50. Un autre motif de citation de *data paper* est illustré par l'exemple du répondant 16, qui a comparé ses propres données aux données décrites dans le *data paper* d'un autre auteur. Le *data paper* a été cité dans la partie « Materials and Methods ». Il semble que le souhait initial de l'auteur était d'utiliser ces données, mais, celles-ci s'étant révélées trop peu précises, il s'en est finalement servi à seul titre de comparaison.

Analyser les données

51. Le répondant 1, dans un troisième article, cite à nouveau le même *data paper* que précédemment, cette fois-ci dans la partie méthodologie de sa publication. Citer le *data paper* lui permet de rappeler le processus de collecte des données, qu'il analyse ensuite dans l'article.

52. Ces différents exemples illustrent des motifs de citation variés, à l'instar des résultats obtenus par Jiao et Darch (2020), allant de la consultation de la méthode de collecte à l'intégration et à la réanalyse des données.

53. Ils montrent également que le *data paper* est utilisé par les chercheurs comme le complément du jeu de données qu'il décrit. Lorsqu'ils utilisent un jeu de données (que ce soit à des fins de compilation, de comparaison ou d'analyse), les auteurs ci-dessus citent systématiquement le *data paper* associé.

54. Dans certains cas, le *data paper* peut aussi être cité pour lui-même. C'est le cas du répondant 4, qui déclare ne pas avoir eu besoin d'accéder au jeu de données. La lecture du *data paper* s'est suffi à elle-même. Parce qu'il fournit des informations sur le contexte de collecte et de traitement des données, le *data paper* peut ainsi être utilisé sans qu'il soit forcément nécessaire de consulter le jeu de données en parallèle.

Le *data paper* comme unité de citation des données ?

55. Les résultats de l'enquête décrits jusqu'ici révèlent des pratiques de citation émergentes, motivées par un *continuum* d'usages axés sur le contenu du *data paper* ainsi que sur celui du jeu de données associé. Dans cette partie, nous développerons deux idées complémentaires, issues des *verbatims* de l'enquête : le *data paper* comme complémentaire de l'article de recherche et sa citation en lieu et place du jeu de données, nous amenant à penser ce nouveau type de publication comme un document situé à l'articulation entre données et article.

Complémentarité entre *data paper* et article de recherche

56. Comme évoqué précédemment (partie « L'importance accordée aux *data papers* »), les *data papers* trouvent leur intérêt dans un système de publication où les articles de recherche sont de plus en plus condensés, ne laissant la possibilité de décrire que succinctement les jeux de données. Dans le champ de l'« archéologie numérique », tel que le nomment Kansa et Kansa (2018), les technologies permettent aujourd'hui de créer des jeux de données de plus en plus complexes, dont le processus d'élaboration est forcément long à détailler et trouve, de fait, difficilement sa place dans les articles de recherche. Les *data papers* viennent ici se positionner à la frontière entre données et articles de recherche. Ils permettent

aux chercheurs de développer les informations contextuelles nécessaires à la compréhension des données tout en délestant les articles de cette partie descriptive :

57. « Parfois, la recherche archéologique crée des jeux de données qui sont difficiles à publier succinctement dans un article descriptif, qui est généralement la façon dont les archéologues publient. Il est très utile de pouvoir mettre les données, clairement exposées et complètes, en un seul endroit, puis d'y faire référence dans l'article interprétatif. »[13]

58. Comme l'explique ce répondant, citer le *data paper* permet de faire référence à la méthodologie de production des données dans l'article, tout en gagnant de la place à l'intérieur de celui-ci. On assiste donc en quelque sorte à une segmentation et une spécialisation des différents types d'articles : les *data papers* pour la description des données, d'un côté ; les articles de recherche pour l'analyse et l'interprétation des données, de l'autre.

59. Si la majorité des répondants à l'enquête jugent positive cette complémentarité entre articles de recherche et *data papers*, deux d'entre eux ont semblé y être moins favorables. Selon eux, les lecteurs ne liront que le *data paper* et pas forcément le ou les article(s) de recherche associé(s) :

60. « Je pense que [les *data papers*] sont néfastes, car ils découragent les lecteurs de consulter les articles originaux et d'obtenir toutes les informations contextuelles. »[14]

61. « Personnellement, je ne les aime pas, car ils détournent l'attention des recherches originales. »[15]

62. Même s'ils ont eux-mêmes publié puis cité un *data paper*, ces chercheurs voient dans la diversification des types d'articles un risque de dispersion des efforts de lecture, qui œuvrerait au détriment de la compréhension du contexte global de la recherche.

Le *data paper*, un point d'entrée vers les données

63. Dans les pratiques des répondants à l'enquête, nous avons également constaté que la citation du *data paper* remplace souvent celle du jeu de données, y compris lorsqu'il y a réutilisation des données. L'analyse de contenu réalisée sur un échantillon de publications (voir plus haut, partie « Motivations à citer un *data paper* du *JOAD* ») montre que, même lorsque les auteurs déclarent avoir cité également le jeu de données (en plus du *data paper*), on ne retrouve pas en réalité sa référence dans la bibliographie. On peut s'interroger sur les raisons de cette contradiction. Il est possible qu'ils ne se soient pas rappelés, au moment de répondre au questionnaire, s'ils

13. « Sometimes archaeological research creates datasets that are difficult to succinctly publish in a descriptive paper, which is usually how archaeologists publish. It is very useful to be able to put the data, clearly set out and complete, in one place, and then refer to it in the interpretive paper. »

14. « I think [*data papers*] are harmful because they discourage readers from consulting the original papers and getting all contextual information. »

15. « I personally don't like them because it takes attention away from the original researches. »

avaient cité ou non le jeu de données. Il est également possible qu'ils assimilent le jeu de données au *data paper* ou simplement que citer le *data paper* revienne pour eux à citer le jeu de données. Cette pratique est peut-être une reproduction du modèle classique de citation, qui consiste à faire principalement référence à des articles et ouvrages. Comme l'a noté Huggett (2018), les archéologues ont l'habitude de citer des documents, non des données structurées. Des entretiens semi-directifs permettraient d'éclairer ce point et ainsi de mieux comprendre la relation entre données et *data paper* dans les pratiques de citation.

64. Si ces réponses invitent à interroger les contours du concept de citation des données, d'autres répondants (2) ont quant à eux clairement exprimé la préférence de citer le *data paper* plutôt que le jeu de données :

65. « Je pense que le *data paper* est un bon point d'entrée vers le jeu de données, et on peut y trouver le lien vers les données. »[16]

66. « J'ai cité le *data paper* qui fait référence au jeu de données. Ainsi, le lecteur peut remonter jusqu'au jeu de données. Après tout, c'est l'un des principaux rôles des *data papers*. »[17]

16. « I think the data paper is a good entry point to the dataset, and the link to the dataset can be found there. »

17. « I cited the data paper that references the dataset. So, the reader can trace his/her way back to the dataset. After all, that is a major role of *data papers* in the first place. »

67. Pour ces auteurs, citer uniquement le *data paper* permet de simplifier le processus de citation (on peut parler de processus en « cascade », de l'article vers les données en passant par le *data paper*). Il permet aussi d'orienter le lecteur vers un document d'un abord plus facile que le fichier de données. C'est l'image du « point d'entrée ». Comme exprimé dans le second *verbatim*, une des fonctions du *data paper* est en effet de faciliter la compréhension du jeu de données grâce à une documentation rédigée et structurée. Le *data paper* deviendrait ainsi l'entité citable du jeu de données.

Conclusion

68. La présente étude a investi la question des pratiques de citation des *data papers* en archéologie à travers le cas d'étude du *Journal of Open Archaeology Data*. En ciblant les auteurs ayant cité un des *data papers* de la revue, l'approche choisie d'une enquête par questionnaire permet de révéler leurs motivations.

69. L'article a montré que la citation des *data papers* en archéologie était le fruit de chercheurs sensibles à l'ouverture des résultats de la science et s'inscrivait dans un contexte d'émergence de ce nouveau type d'articles. Les résultats de l'enquête, arrêtée à la date du 17 août 2021, montrent également que la citation des *data papers* traduit bien plus d'usages que la seule réutilisation des données, celle-ci pouvant elle-même prendre différentes formes (de l'analyse à la comparaison ou la compilation

des données avec d'autres jeux de données). Le *data paper* est cependant rarement cité pour lui-même : sa citation renvoie le plus souvent aux données sous-jacentes qu'il documente. Pétris par des pratiques de citation centrées sur les documents de la littérature scientifique (articles, ouvrages, rapports…), les chercheurs reproduisent ici un mécanisme qui relève de l'habitude, préférant citer le *data paper* plutôt que le jeu de données. Grâce à sa fonction de contextualisation, le *data paper* constitue un point d'accès qui permet de se familiariser avec les données et de juger de leur pertinence pour l'usage que l'on souhaite en faire. Ce faisant, il allège également l'article de recherche, dans lequel la description du processus de collecte des données se limite alors à la citation du *data paper*.

70. Ces résultats nous invitent donc à prolonger notre recherche en étudiant comment la citation des *data papers* s'insère plus largement dans les pratiques de citation des chercheurs en archéologie, nous permettant ainsi de contribuer, sous un prisme disciplinaire, à une meilleure compréhension de la présence des *data papers* dans le système de la communication scientifique.

Les données mobilisées dans le chapitre sont disponibles dans l'entrepôt Nakala (Huma-Num). Elles sont réutilisables selon les termes de la licence CC-BY-NC-4.0. Le jeu de données comprend la liste des publications ayant cité un des *data papers* du *JOAD*, le questionnaire de l'enquête, ainsi que les réponses des 19 participants.

Pour citer le jeu de données : Rebouillat, Violaine. 2022. « Étude des formes et logiques de citation des *data papers* du Journal of Open Archaeology Data » [Survey data] Nakala. https://doi.org/10.34847/nkl.e35572h8

Le *Journal of Open Humanities Data* (*JOHD*) : enjeux et défis dans la publication de *data papers* pour les sciences humaines et sociales (SHS)

Paola Marongiu, Nilo Pedrazzini, Marton Ribary et Barbara McGillivray

Introduction

1. La publication de *data papers* pour les Sciences humaines et sociales (SHS) est une activité qui compte déjà une dizaine d'années, mais qui est néanmoins très actuelle. L'idée de pouvoir publier des articles consacrés exclusivement à la description des données utilisées dans un cadre de recherche, et non à la recherche elle-même et à ses résultats, est née, au moins pour les sciences dures, dans les années 1950. Des années plus tard, après le croisement des SHS avec des méthodes et des outils scientifiques, il est devenu courant de s'interroger sur la pertinence, voire la nécessité, de mettre en évidence le travail considérable effectué pour élaborer les données utilisées dans la recherche en Sciences humaines et sociales. Cette situation a entraîné des questions sur la définition même de « donnée » en SHS et a révélé une grande variation au sein des différentes disciplines quant à la nature des objets qu'elles étudient et qu'elles produisent. Ce type de questions et de besoins a conduit à la naissance du premier *data journal* en SHS, en 2012 ; celui-ci était consacré au domaine de l'archéologie. Dès lors, les débats sur la production et l'exploitation des données en SHS se sont intensifiés avec une acuité particulière donnée à l'importance des valeurs de la science ouverte, du partage et de la réutilisation des données. C'est précisément dans ce contexte que s'inscrit le *Journal of Open Humanities Data* (*JOHD*), qui cherche à répondre au besoin de reconnaissance du travail en SHS non seulement du point de vue des résultats scientifiques, mais aussi du point de vue des étapes en amont. Concrètement, cela inclut la définition et l'identification des données utilisées dans une recherche, la création d'un jeu de données bien structuré et lisible par une machine, et finalement le traitement et l'analyse des données au moyen de scripts ou d'outils spécifiques. Toutes ces opérations nécessitent une expertise de la part des chercheurs ainsi qu'un investissement considérable en matière de temps. Dans ce chapitre, nous essaierons de faire le point sur les définitions possibles du *data paper* et sur les objectifs de ce type de publication. Nous analyserons ensuite l'histoire de la publication de *data papers* au fil du temps, d'abord dans le monde de la recherche en général, et plus spécifiquement pour les SHS. Dans ce contexte, nous présenterons la revue *JOHD*, en présentant sa structure, ses objectifs et ses activités éditoriales entre autres. Nous nous intéresserons par la suite à l'impact que peut avoir la publication de *data papers* sur la réutilisation des données. Enfin, nous

proposerons notre modèle pyramidal dans lequel les *data papers*, les entrepôts ouverts, les entrepôts de projet et les articles traditionnels participent à la diffusion de la recherche en accès ouvert.

Data papers : définition et contexte

Définition et rôle d'un *data paper*

2. La question de la définition du *data paper* ayant été abordée dans ce volume par Victoria le Founer et Joachim Schöpfel, nous nous concentrerons sur la définition propre au *Journal of Open Humanities Data* (*JOHD*).

3. Pour le *JOHD*, « un *data paper* est une publication destinée à faire connaître à d'autres chercheurs et chercheuses des données qui pourraient leur être utiles. En tant que tel, il décrit les méthodes utilisées pour créer l'ensemble de données, sa structure, son potentiel de réutilisation et fournit un lien vers son emplacement dans un entrepôt[1] ». Il est également souligné que l'objectif d'un *data paper* est différent de celui d'un article de recherche, qui est plutôt censé illustrer le processus d'enquête scientifique. Les deux ont en effet des rôles complémentaires.

4. Les valeurs de la science ouverte jouent un rôle fondamental dans l'encouragement à la publication des données et des codes sources en accès ouvert. La notion de science ouverte est précisée par l'*Open Knowledge Foundation* : « Ouvert, signifie que tout le monde peut librement accéder, utiliser, modifier et partager dans n'importe quel but (sous réserve, tout au plus, d'exigences qui préservent la provenance et l'ouverture)[2]. » Au cours des dernières années, les valeurs de science ouverte ont été promues par diverses initiatives, institutions et divers projets. Dans de nombreux pays, les fonds de financement public (par exemple, le *Fonds National suisse*) appliquent des contraintes concernant la publication des résultats de recherche des projets qu'ils financent. Souvent, ceux-ci doivent être disponibles en accès ouvert. De plus, depuis 2018, la *cOAlition S* s'engage, à travers le *Plan S*, pour la publication en accès ouvert des résultats de recherche de tous les projets financés par les institutions de financement (publiques et privées) qui ont décidé de rejoindre la coalition. Parmi ces institutions, nous pouvons mentionner le *UK Research and Innovation*, *The Research Council of Norway* et le *National Science Centre – Poland* entre autres[3]. D'autres initiatives sont représentées par la *Declaration on Research Assessment (DORA)*[4], établie en 2012, et l'établissement des principes *FAIR*[5] en 2016. La *DORA* s'engage pour l'amélioration des méthodes d'évaluation des résultats

1. https://openhumanitiesdata.metajnl.com/about/#q9

2. http://opendefinition.org/

3. Les projets partenaires sont disponibles à l'adresse https://www.coalition-s.org/organisations/. Pour le *Plan S* et la *cOAlition S*, nous renvoyons à l'adresse suivante : https://www.coalition-s.org/why-plan-s/.

4. https://sfdora.org/about-dora/

5. https://www.go-fair.org/fair-principles/

de recherche, dans toutes les disciplines, et pour l'assouplissement des inégalités au sein du système académique. Quant aux principes *FAIR*, leur objectif est de donner des lignes directrices pour la gestion du flux de travail de la recherche. Cette notion est détaillée dans l'introduction du présent ouvrage. Dans le domaine des SHS, des initiatives, telles que *SSHOC*[6], *LIBER*[7], *CLARIN*[8], entre autres, contribuent à promouvoir ces valeurs et à offrir aux chercheurs et chercheuses un réseau de soutien et les moyens de poursuivre dans cette voie. En effet, la communauté scientifique se rend progressivement compte des avantages que la recherche peut tirer de la mise en œuvre des bonnes pratiques d'accès ouvert. Ces avantages se recoupent avec les objectifs de la publication de *data papers* : ils garantissent une meilleure transparence concernant les données, les méthodes et les outils utilisés dans une recherche ; ils donnent la possibilité d'obtenir un avis et/ou une contribution externes de la part d'autres chercheurs, ce qui peut participer à améliorer la qualité des données ainsi que des résultats ; ils offrent de nouvelles options pour la diffusion de la recherche individuelle, élargissant le public cible ; indirectement, ils encouragent les chercheurs et les chercheuses à fournir des jeux de données bien organisés, avec une structure interne claire, et accompagnés d'une riche documentation, pour en faciliter l'utilisation externe.

6. https://sshopencloud.eu/about-sshoc
7. https://libereurope.eu/about-us/
8. https://www.clarin.eu/content/about-clarin

La publication de *data papers*

5. La publication de *data papers* a connu un grand essor ces derniers temps. Garcia-Garcia *et al.* (2015) en ont reconstitué l'histoire de la publication depuis le début en 1956, avec le lancement de la première revue de données (ou *data journal*), le *Journal of chemical and engineering data*, jusqu'en 2015. Cette étude met en évidence un démarrage tardif des revues de données dans les sciences humaines et sociales (SHS) par rapport aux sciences dures. Le premier *data journal* consacré aux SHS a été le *Journal of Open Archaeology Data* lancé en 2012. Un nouvel éclairage sur ce thème a été donné par Schöpfel *et al.* (2019) qui ont mis en lumière l'étude réalisée par Candela *et al.* (2015) sur le nombre de revues de données et les domaines d'intérêt couverts par celles-ci. Leurs résultats montrent que le nombre de revues de données n'a pas énormément augmenté depuis l'enquête de 2015 (20 en 2015 et 28 en 2019). Néanmoins, le nombre de *data papers* publiés est passé de 846 en 2013 à 11 500 en 2019. Cependant, les derniers résultats élaborés par Walters (2020) révèlent que seulement cinq revues de données en 2020 étaient consacrées aux SHS. Il s'agit du *Journal of Open Archaeology data* (2012), du *Journal of Open Psychology data* (2013), du *Journal of Open Humanities data* (2015), du *Research Data in the Humanities and Social Sciences* (2016), et du *Internet Archaeology* (1996) qui publie plutôt des représentations numériques d'artefacts. La raison de cet impact plus tardif des *data papers* sur les disciplines en SHS peut s'expliquer par un certain nombre de raisons.

6. Par rapport aux sciences dures, le domaine des SHS est assez étendu et hétérogène. La linguistique, les sciences sociales, les études cinématographiques, la philologie, l'histoire, les études sur les jeux, la philosophie ne sont que quelques-uns des domaines couverts par le terme « humanités ». Non seulement, ces disciplines s'intéressent à des objets d'étude complètement différents, mais les données qu'elles produisent ne sont souvent pas comparables. En effet, l'un des problèmes principaux des SHS est la définition de la donnée elle-même. Les collections de textes (corpus), les enregistrements de dialogues, les vidéos, les éditions numériques représentent des types différents de données dans leurs disciplines respectives. Il est assez intuitif que l'analyse d'un dialogue enregistré afin de détecter différentes variations phonétiques ne soit pas comparable, en matière de raisonnement et d'approche, à la publication d'une édition numérique d'un manuscrit inédit. Pourtant, tous deux sont considérés comme des données dans les deux disciplines, respectivement la phonologie et la philologie numérique. Ce manque d'homogénéité a un impact sur la manière dont la plupart des chercheurs et des chercheuses dans les humanités conçoivent leur travail et leurs données. En particulier, cette situation de départ entraîne l'absence d'un ensemble de critères, communs à toutes les disciplines en SHS, permettant de déterminer si un jeu de données est prêt à être publié.

7. Certaines de ces disciplines travaillent avec des ensembles de données fermés (c'est-à-dire des objets d'étude qui ne peuvent plus être étendus, comme les documents historiques, par opposition, par exemple, aux données numériques). Les chercheurs et les chercheuses dans ces domaines ne considèrent souvent pas leurs données comme prêtes à être décrites formellement tant qu'elles n'ont pas conduit à certains résultats de recherche. À titre d'exemple, un chercheur ou une chercheuse qui travaille avec des sources couvrant toute la période de la révolution industrielle pourrait supposer qu'un ensemble de données produites à partir d'un petit échantillon de ces sources (par exemple, couvrant seulement quelques années de la période concernée) n'ait aucun intérêt pour la communauté scientifique. Or, les *data papers* détournent l'attention des objectifs d'un projet de recherche ou d'une question spécifique vers la façon dont les données ont été collectées et présentées, et vers la possibilité d'application de la même méthode à des sources de données similaires. Par conséquent, le chercheur ou la chercheuse qui travaille sur la révolution industrielle peut donc décider de décrire ce petit échantillon de données dans un *data paper* si la même méthode est appliquée au reste des sources, et surtout si le processus de création de l'échantillon a été objet d'une planification minutieuse. La publication d'un *data paper* sera ainsi profitable à la communauté, indépendamment du fait que l'ensemble des données qu'il décrit est déjà représentatif d'un aspect de la révolution industrielle. Dans ce contexte, les chercheurs et les chercheuses en SHS qui travaillent avec des ensembles de données ouvertes, tels que des objets numériques (par exemple, des données X ou des médias d'information), ne sont en fin de compte pas si différents des historiens qui

étudient la révolution industrielle : la collecte, l'organisation et la description des données sont des processus qui méritent d'être reconnus en tant que tels. En même temps, le chercheur devrait pouvoir donner un aperçu de la manière dont ce processus peut bénéficier à d'autres au sein de la communauté scientifique.

8. Bien que cela soit en train de changer, l'idée de partager un ensemble de données en tant que résultats de recherche n'est pas encore suffisamment répandue, car les données ne sont souvent pas reconnues en tant que telles. Par exemple, les revues de données ne figurent pas souvent dans les listes de revues reconnues pour l'obtention de crédits académiques dans de nombreux pays. En outre, lorsqu'il s'agit d'évaluer les résultats de la recherche, les articles de recherche traditionnels font souvent l'objet d'une évaluation, alors que les *data papers* ne sont pas pris en compte dans la même mesure. Par exemple, dans le *REF* (*Research Excellence Framework*), en Angleterre, qui s'occupe d'évaluer les résultats de recherche des universités anglaises, les résultats soumis pour l'évaluation sont pour la majorité encore des articles publiés dans des revues traditionnelles.

9. Par conséquent, on constate que le flux de travail d'un chercheur ou d'une chercheuse en SHS n'inclut pas en général la publication des données. Le fait que le flux de travail ne prévoit pas d'étapes pour l'organisation des données en vue de leur publication fait que la documentation des jeux de données est souvent insuffisante, et que leur modalité et leur potentialité de réutilisation restent souvent obscures pour un public externe. De plus, une charge de travail supplémentaire est nécessaire pour les rendre publiables. À cela s'ajoute la rareté ou le manque de fonds pour supporter les coûts de stockage ou les frais de publication des articles (APC). C'est souvent le fonds qui finance le projet de recherche qui couvre ce genre de dépenses, mais si un chercheur ou une chercheuse ne peut pas compter sur un financement externe, ce qui est plus courant en SHS que dans les sciences dures, il ne sera pas possible de prendre en charge les coûts de stockage et publication en accès ouvert avec APC.

10. La méconnaissance des infrastructures d'archivage et de stockage des données représente parfois aussi un obstacle. Cette situation peut être due à plusieurs facteurs. Par exemple, souvent, il n'y a pas assez de formations au sein des universités visant à acquérir les outils et les connaissances relatifs à la diffusion en accès ouvert des résultats de la recherche. On constate également que de nombreuses infrastructures dédiées au stockage de données ne sont souvent pas adaptées pour le stockage de données produites par les SHS. Pour faire face à ce type de problème, par exemple, le projet *SSHOC*[9] et ses partenaires visent à mettre en place des réseaux intégrés d'infrastructures de données interconnectées.

9. Le *Social Sciences & Humanities Open Cloud* (https://sshopencloud.eu/) est un projet qui compte 20 organisations partenaires et qui vise à la création d'une *European open science Cloud* (https://eosc-portal.eu/about/eosc) pour l'hébergement des données de recherche en SHS. Le projet a été financé par le programme européen Horizon 2020.

11. Malgré ces difficultés, on remarque une tendance positive ces dernières années concernant le partage de données en SHS. De nombreux facteurs ont contribué à ces résultats. Nous avons déjà mentionné les projets, les conférences et les institutions de financement qui s'engagent dans la promotion des valeurs de la science ouverte[10]. Un autre facteur important est la numérisation progressive de collections physiques, par exemple de livres ou d'œuvres d'art. En général, on peut compter en SHS sur un nombre majeur de ressources numériques, telles que des langages de programmation performants, ainsi que des outils pour créer, lire ou analyser différents types de jeux de données. Enfin, l'émergence récente de la science des données en tant que discipline a permis une nouvelle prise de conscience du rôle central des données dans les recherches en SHS. En effet, elle a également mis en évidence la nécessité de rendre les données librement accessibles et de leur assurer un stockage à long terme, comme c'est déjà le cas pour les sciences dures.

Les *data papers* aujourd'hui : la pyramide de l'accès ouvert

12. Les *data papers* font partie d'une stratégie de recherche et de diffusion à plusieurs niveaux que nous proposons d'appeler « la pyramide de l'accès ouvert » (figure 1). Dans cette section, nous décrirons les quatre niveaux de la pyramide et le rôle que chacun d'entre eux joue pour rendre la recherche totalement transparente, reproductible et efficace : (1) l'entrepôt de projet, (2) l'entrepôt de données, (3) le *data paper*, et (4) l'article de recherche. Ces quatre niveaux seront expliqués à partir d'un projet de recherche en humanités numériques, dont le *data paper* a été publié dans *JOHD* (Ribary, 2020a), avec des liens vers les trois autres niveaux. Nous soutenons que, contrairement au modèle traditionnel qui se concentre presque exclusivement sur l'article de recherche, celui de la pyramide de l'accès ouvert accorde le crédit nécessaire à la complexité du processus de recherche et de diffusion. De plus, ce modèle souligne le fait que l'article de recherche au sommet de la pyramide ne présente qu'une seule interprétation possible des données collectées, et qu'elles sont toujours ouvertes à des analyses alternatives.

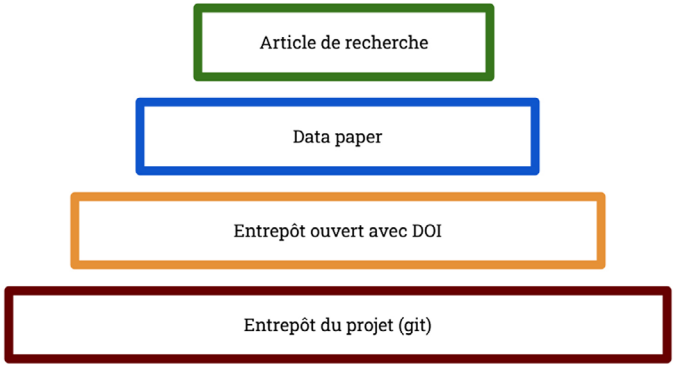


Figure 1. La pyramide de l'accès ouvert

Crédit : Ribary (2021), Surrey Open Research and Transparency Showcase 2021

10. En plus de *SSHOC*, *CLARIN* et *LIBER* déjà mentionnés, nous pouvons ajouter à la liste l'initiative *Research Data Alliance (RDA)* (https://www.rd-alliance.org/) et tous les événements organisés par la *RDA* ; le colloque *#dhnord2021* (https://www.meshs.fr/page/dhnord2021) ; le projet *EOSC-hub* (https://www.eosc-hub.eu/about-us) ; le projet *Open Data Literacy (ODL)* (http://odl.ischool.uw.edu/) ; le portail *data.europa.eu* (https://data.europa.eu/en).

13. Au bas de la pyramide se trouve l'entrepôt de projet qui contient tous les fichiers de ressources et les fichiers intermédiaires, les documents d'accompagnement et les scripts utilisés pour le traitement des données. L'entrepôt de projet est le lieu de travail public et de stockage du chercheur, où le contrôle des versions (via git, par exemple) garantit qu'aucun travail n'est perdu, et où l'on peut revenir aux versions précédentes si une direction de recherche s'avère être une impasse. La documentation de l'entrepôt du projet doit être suffisamment détaillée pour rendre la recherche transparente et reproductible, mais elle ne doit pas ralentir inutilement le projet. Le projet que nous utilisons comme exemple dans cette section s'inscrit dans le domaine des humanités numériques et il porte sur la structure et le langage du recueil de lois romaines de l'empereur Justinien connu sous le nom de *Digeste* (533 EC) (Ribary, 2020c). Dans ce cadre, un organigramme de traitement développé de manière incrémentielle (figure 2) nous a aidés à suivre les fichiers intermédiaires. En fait,

Figure 2. Un organigramme de traitement correspondant à l'entrepôt de projet de Ribary (2020c)

Crédit : Ribary (2020c), pyDigest : A GitLab repository of scripts, files and documentation

l'organigramme a servi de table des matières visuelle du projet. Il s'est avéré incroyablement utile lorsqu'un utilisateur a repéré un désalignement dans la base de données publiée (Ribary, 2020b) à la suite d'une erreur de codage dans un script. L'organigramme et la documentation interne des scripts ont permis de localiser l'erreur de codage assez rapidement. Il a également été possible d'exécuter à nouveau les étapes de traitement et de créer une version corrigée de la base de données.

14. Le deuxième niveau de la pyramide est représenté par l'entrepôt où les données ont été déposées. Les données structurées y sont présentées avec une documentation allégée et les instructions nécessaires à leur réutilisation. Le projet en question (Ribary, 2020b) a utilisé l'entrepôt Figshare[11] qui était également l'emplacement de la base de données révisée. Figshare, comme d'autres entrepôts, permet de publier des mises à jour sans supprimer les versions précédentes. L'entrepôt fournit un DOI qui est marqué avec le numéro de version. Dans notre cas, la base de données SQLite publiée était également accompagnée d'exemples de requêtes, afin de ne pas limiter l'accès aux connaisseurs du langage SQL. C'est en effet un utilisateur novice en SQL qui a repéré un mauvais alignement des données et nous l'a signalé.

11. Figshare est un entrepôt à accès ouvert. Les fichiers déposés pour ce projet sont : la base de données relationnelle du *Digeste* ; un fichier en format txt qui contient des exemples de requêtes en langage SQL qui peuvent être utilisées sur la base de données ; un fichier README qui donne une description du projet et des instructions pour lancer les requêtes sur la base de données ; un organigramme qui montre le schéma de la base de données.

15. Le troisième niveau de la pyramide est le *data paper*, qui indique et décrit les données publiées dans l'entrepôt. Le *data paper* attire l'attention sur la ressource et souligne son potentiel de réutilisation. Il donne l'occasion de décrire le contexte historique et méthodologique du projet et de fournir un résumé narratif de la production des données. Le *data paper* indique également des pistes de recherche possibles qui prennent les données comme point de départ. Alors que les données ont été créées pour répondre à une question de recherche spécifique, qui est élaborée dans un article de recherche traditionnel, le *data paper* ouvre idéalement les données à des projets bien au-delà de cette portée.

16. L'article de recherche traditionnel se trouve au sommet de la pyramide. Il s'agit du résultat le plus prisé de la recherche universitaire qui, trop souvent, adopte une vision étroite de la recherche et rejette le travail que les chercheurs et les chercheuses ont pu publier dans les trois premiers niveaux. Bien que la tendance soit en train de changer, la publication d'un article de recherche reste un processus notoirement lent et qui, dans de nombreux cas, produit un article à accès payant. Ces deux points créent une friction évidente avec les trois couches précédentes, où les résultats sont mis à la disposition de tous dans des délais plus courts ou immédiatement. Afin d'éviter ces frictions, il est possible d'envisager la publication en accès ouvert dans un journal en ligne. Le projet cité en exemple a choisi cette voie en ciblant le *JOHD* pour le *data paper* (Ribary, 2020a) et une revue académique similaire pour l'article de recherche correspondant

(Ribary et McGillivray, 2020). Si les quatre niveaux de la pyramide sont publiés dans une succession relativement rapide, et s'ils adhèrent aux principes de science ouverte, les chercheurs et les chercheuses pourront donner à leur travail la meilleure chance d'avoir un impact.

Le *Journal of Open Humanities Data* (*JOHD*)

Présentation de la revue

17. Le *Journal of Open Humanities Data* (*JOHD*) a été lancé en 2015 dans le but de promouvoir les valeurs de partage et de réutilisation des données dans le vaste domaine des SHS. *JOHD* est l'une des méta-revues[12] publiées par Ubiquity Press[13], éditeur d'articles scientifiques en accès ouvert. Leur objectif est d'encourager le partage de ces ressources selon les bonnes pratiques de la science ouverte. En retour, les revues offrent aux auteurs la possibilité de publier un article scientifique, d'obtenir des citations, de faire connaître les données et d'accroître la réutilisation de leurs ressources. Cet aspect est particulièrement important, car, dans le cadre du système actuel axé sur les publications, les chercheurs et les chercheuses qui produisent du code ou travaillent sur les données ne bénéficient pas d'une reconnaissance académique appropriée.

18. Le *JOHD* publie des articles axés sur les données et vise à jouer un rôle clé dans le développement d'une communauté de chercheurs et de chercheuses en SHS qui partagent leurs données. Les articles publiés dans le *JOHD* sont en anglais jusqu'à aujourd'hui. Ce choix est dû à deux raisons : la première est liée à l'identité de la maison d'édition, Ubiquity Press, qui est née en Angleterre, travaille principalement en anglais ; la seconde est que l'utilisation de la langue anglaise permettait de mieux atteindre le public de la revue qui, au moins au début, était majoritairement anglophone. Une condition fondamentale pour la publication dans le *JOHD* est d'avoir déposé le jeu de données dans un entrepôt en accès ouvert. Parmi les entrepôts recommandés, on peut mentionner le *JOHD Dataverse*[14] conçu spécifiquement pour les *data papers* publiés par la revue.

19. Au cours des deux dernières années, le *JOHD* s'est considérablement développé. Les raisons de cette évolution ont plusieurs origines. Le rédacteur en chef a consacré beaucoup d'attention à la promotion de la revue et à l'amélioration de son profil au sein de la communauté universitaire et des bibliothèques. Grâce à l'élargissement du comité de rédaction et à l'engagement d'activités scientifiques et de promotion du *JOHD* dans de nouveaux pays, nous avons pu élargir la portée

12. Il s'agit de revues qui s'intéressent à la publication d'articles décrivant des outils de recherche ainsi que le processus qui se cache derrière les données (par exemple la conception et l'utilisation de logiciels ou de matériel informatique, et la construction d'ensembles de données).

13. https://www.ubiquitypress.com/

14. https://dataverse.harvard.edu/dataverse/*JOHD*.

disciplinaire et géographique de la revue. L'équipe s'est également agrandie avec un certain nombre de fonctions différentes occupées par des étudiants et des chercheurs en début de carrière. Les tâches éditoriales sont partagées entre les assistants éditoriaux et la gestion des médias sociaux repose sur la contribution constante et régulière du rédacteur des médias sociaux de la revue. En outre, le rédacteur en chef et l'équipe ont mené un certain nombre d'activités où la revue a joué un rôle de leader dans la publication de données pour les SHS, notamment en organisant des événements scientifiques[15] et en donnant des interviews[16], auxquels s'ajoutent la participation aux colloques[17] et aux panels de discussion[18], la présentation de posters[19] dans des événements internationaux[20] et la publication d'articles de blog[21] et de publications académiques (Engelhardt *et al.*, 2022). Enfin, une stratégie éditoriale ciblée a permis de lancer des collections d'articles sur des thèmes ou des sous-disciplines spécifiques[22], ce qui a permis à la revue d'atteindre des communautés scientifiques dans des secteurs de recherche spécialisés.

20. La stratégie de médias sociaux adoptée par le *JOHD* a contribué à l'accroissement du volume d'audience de la revue, principalement en augmentant le nombre de followers sur ses pages. Pour ce faire, nous avons développé un ensemble de hashtags susceptibles d'orienter le public sur le type d'informations que nous voulions diffuser. En utilisant les hashtags *#johdagenda* et *#johdsuggestions*, nous nous engageons dans la

15. Y compris le premier (https://youtu.be/KJN7X_nizqI) et le deuxième *Open Humanities Data Forum* (https://youtu.be/OoZgyK6TdVU), organisés en concomitance avec la conférence plénière de la *Research Data Alliance (RDA)* en 2021.

16. Y compris l'interview avec Shalhavit-Simcha Cohen pour la communauté PosiFest (https://www.facebook.com/posifest/posts/232103365001120), et avec Dennis Relojo-Howell pour *PsychReg* (https://www.youtube.com/watch?v=q6-mzNqOpGc).

17. Y compris la conférence « *Data citation for the Humanities and Social Sciences : a special case?* » de Barbara McGillivray, Nicolas Larrousse et Daan Broeder au colloque *LIBER* en 2021 (https://www.youtube.com/watch?v=nKgUT_3hkDg&list=PLHA3lUmrYM3u-Rs-L52lrzmLcjRKr3RJq&index=10) et le lightning talk « *Open data and the digital humanities in the time of COVID-19* » de Mandy Wigdorowitz, Sahba Besharati, et Barbara McGillivray au colloque *DHASA* en 2021 (https://dh2021.digitalhumanities.org.za/schedule/).

18. Y compris le panel de discussion « *FAIR Data-Citation for Social Sciences and Humanities* » (https://www.eosc-hub.eu/events/realising-european-open-science-cloud/fair-data-citation-ssh), la table ronde « *Round Table of Experts on Data Citation* » (https://www.sshopencloud.eu/events/round-table-experts-data-citation) et la table ronde « *Le data paper : une nouvelle forme de publication en SHS* » (https://www.meshs.fr/page/data_paper_une_nouvelle_forme_de_publication_scientifique_en_shs).

19. Y compris le poster « *Publish peer-reviewed Humanities data papers in JOHD* » présenté au colloque *eResearch Australasia* en 2021 par Toby Burrows, Ingrid Mason et Barbara McGIllivray (https://conference.eresearch.edu.au/) et le poster « *What does it mean to publish historical sources today?* » de Fernanda Olival, Helena Freire Cameron, Renata Vieira, Ivo Santos et Barbara McGillivray présenté au *Linked Pasts VII Symposium* en 2021 (https://www.ghentcdh.ugent.be/linked-pasts-vii-symposium).

20. Y compris le panel de discussion « *FAIR Data-Citation for Social Sciences and Humanities* » (https://www.eosc-hub.eu/events/realising-european-open-science-cloud/fair-data-citation-ssh), la table ronde « *Round Table of Experts on Data Citation* » (https://www.sshopencloud.eu/events/round-table-experts-data-citation) et la table ronde « *Le data paper : une nouvelle forme de publication en SHS* » (https://www.meshs.fr/page/data_paper_une_nouvelle_forme_de_publication_scientifique_en_shs).

21. Y compris « *Write an Open Data Paper : an invitation to JOHD* » de Gabriel Bodard et Barbara McGillivray (https://blog.stoa.org/archives/3877) et « *Publishing the Mapping Manuscript Migrations Data* » de Toby Burrows (https://blog.mappingmanuscriptmigrations.org/blog/).

22. https://openhumanitiesdata.metajnl.com/collections

conversation sur les données ouvertes et les humanités numériques sur les médias sociaux, en retweetant les nouvelles et les événements annoncés par les associations et les projets actifs dans ces domaines ; avec les hashtags *#johdguides* et *#johdpapers*, nous décrivons la politique et le processus éditorial de la revue et nous annonçons les nouvelles publications ; enfin, grâce aux hashtags *#johdnews* et *#johdCfP*, nous tenons notre public au courant des activités de l'équipe et des nouveaux appels à contributions. L'activité sur X vise également à sensibiliser la communauté aux valeurs de la science ouverte et du partage des données, et cela, notamment à travers la campagne *#showmeyourdata*. Dans le cadre de cette initiative, nous invitons les auteurs qui ont publié un article avec le *JOHD* à publier un tweet montrant une image du jeu de données qu'ils ont décrit dans leurs articles. La plupart d'entre eux ont publié une capture d'écran de leur ensemble de données, ce qui a montré à quel point la définition de « données » peut être variée dans le domaine des SHS : nous avons reçu des images de fichiers dans des entrepôts, de réseaux, de codes, de cartes, de feuilles de calcul[23].

21. Tous ces efforts ont contribué à positionner la revue à l'avant-garde de la publication de *data papers* en SHS et ont conduit à un nombre croissant de publications, ce qui a lui-même contribué à accroître sa notoriété au sein de la communauté.

23. Tous nos tweets pour cette campagne et les retweets de nos auteurs peuvent être retrouvés sur X en utilisant le hashtag *#showmeyourdata*.

Structure d'un *data paper* et d'un *research paper*

22. Le *JOHD* publie deux types d'articles : des articles courts sur les données (*data papers*) et des articles de recherche plus longs (*research papers*). Les deux sont axés sur la description et/ou le processus de création et d'exploitation des données. La publication d'un *data paper* ou d'un article de recherche avec *JOHD* est expliquée en détail dans la section correspondante de notre site web[24].

23. Les *data papers* sont des contributions plutôt courtes (1 000 mots environ), qui visent à décrire de manière concise un ensemble de données, les méthodes utilisées pour le créer et son potentiel de réutilisation. Ils prévoient généralement une première section sur le contexte dans lequel l'ensemble de données a été produit (une thèse, un projet de recherche, etc.). La section suivante décrit les méthodes utilisées pour recueillir et traiter les données. Il s'agit de la description détaillée de la procédure, le cas échéant, la stratégie d'échantillonnage utilisée, les méthodes appliquées afin d'effectuer un contrôle de qualité sur les données. Ensuite, les auteurs décrivent le jeu de données en précisant le nom du/des fichiers déposés, leur format d'encodage, les dates de création ; les noms des créateurs, la ou les langues utilisées dans le jeu de données, la licence ouverte utilisée pour l'entrepôt, le nom de l'entrepôt ouvert, la date à laquelle l'ensemble de données a été publié dans l'entrepôt. La dernière section est consacrée à la

24. https://openhumanitiesdata.metajnl.com/about/submissions/

description des possibilités de réutilisation des jeux de données par d'autres chercheurs et chercheuses et pour d'autres fins de recherche[25].

24. Les articles de recherche sont des textes plus longs, de 3 000 à 5 000 mots. Ils décrivent les méthodes, les défis et les limites du processus de création, d'analyse et d'utilisation d'un ensemble de données dans la recherche en SHS[26].

Étapes de publication d'un *data paper* dans le JOHD ?

25. Les directives de publication d'un *data paper* ont été préparées en consultation avec les auteurs. Elles concernent le type de données pour lesquelles un *data paper* peut être publié par le *JOHD* et les étapes à suivre pour que les données soient ouvertes. Comme mentionné ci-dessus, la première condition pour publier avec le *JOHD* est d'avoir déposé les données dans un entrepôt en accès ouvert. Il peut s'agir d'un entrepôt suggéré par le *JOHD* ou d'un autre entrepôt au choix de l'auteur et approuvé par la revue. Les entrepôts suggérés aux auteurs sont JOHD Dataverse, Figshare, Zenodo, SND et DANS. Dans la section « Soumissions » du site du *JOHD*[27], les auteurs peuvent trouver des onglets consacrés à la description de chaque entrepôt, afin qu'ils puissent se renseigner brièvement sur leurs caractéristiques principales et faire un choix plus éclairé. L'accessibilité de l'ensemble des données doit être garantie par une déclaration qui vise à illustrer la manière dont les lecteurs et les autres utilisateurs en général peuvent accéder aux données. Cette déclaration comprend également le DOI fourni par l'entrepôt public dans lequel les données ont été stockées.

26. Les soumissions peuvent se faire de différentes façons. La revue accepte régulièrement les soumissions spontanées d'articles. En plus de cette option, il existe la possibilité de soumettre un article pour un numéro thématique. Ainsi, *JOHD* a publié 3 numéros thématiques : « *Humanities Data in the time of COVID-19*[28] » qui accueille des études sur les mécanismes et les conséquences de la pandémie de COVID-19 du point de vue des SHS ; « *Language documentation collections : assessment and recognition*[29] » pour des articles sur la documentation des langues et des pratiques linguistiques menacées ; « *Computational humanities research*

25. Template pour les *data papers* en Word (https://s3-eu-west-1.amazonaws.com/ubiquity-partner-network/up/journal/johd/short%20data%20paper%20template.docx) et en LaTeX (https://www.overleaf.com/latex/templates/johd-data-paper-template/mqcypcbntsds).

26. Template pour les articles de recherche en Word (https://s3-eu-west-1.amazonaws.com/ubiquity-partner-network/up/journal/johd/Full%20length%20research%20paper%20template.docx) et en LaTeX (https://www.overleaf.com/latex/templates/johd-research-paper-template/qcthhdwhrhps).

27. https://openhumanitiesdata.metajnl.com/about/#repo

28. https://openhumanitiesdata.metajnl.com/collections/humanities-data-in-the-time-of-covid-19

29. https://openhumanitiesdata.metajnl.com/collections/language-documentation-collections-assessment-and-recognition

data[30] » qui a été lancé à l'occasion du *First Computational Humanities Research Workshop (CHR2020)* et qui s'intéresse aux défis posés par l'application des méthodes computationnelles aux données produites en SHS.

27. L'évaluation d'un *data paper* commence par l'ensemble des données et la manière dont elles sont déposées et documentées dans l'entrepôt. La validité des données n'est pas évaluée en soi, ce qui est important est qu'elles soient librement accessibles. L'évaluation par les pairs évaluera si l'entrepôt est adéquat et s'il dispose d'un modèle de durabilité ; si les données sont déposées sous une licence ouverte ; si l'ensemble de données est compréhensible par d'autres utilisateurs et s'il est complet. Souvent, la difficulté d'utiliser les jeux de données d'autrui tient au fait qu'ils ont été créés pour un usage interne. La conscience de la possibilité d'un usage externe du jeu de données encourage ses créateurs à l'organiser et le documenter de manière claire, en étiquetant tous les éléments qui seraient autrement obscurs pour un utilisateur externe. Si le jeu de données est composé d'informations personnelles provenant de sujets humains, il doit être anonymisé et les créateurs doivent obtenir le consentement éclairé de tous les sujets. L'éthique est un autre élément fondamental pour les jeux de données impliquant des participants humains : les chercheurs et les chercheuses doivent prouver qu'ils ont collecté les données conformément à la déclaration d'Helsinki[31] et que l'étude a été approuvée par un comité d'éthique. Enfin, le format dans lequel les données sont stockées doit être ouvert et gratuit.

28. La revue a publié 38 articles entre septembre 2015 et août 2021. La majorité (75,7 %) est représentée par des *data papers* (29). Les neuf autres sont des articles de recherche.

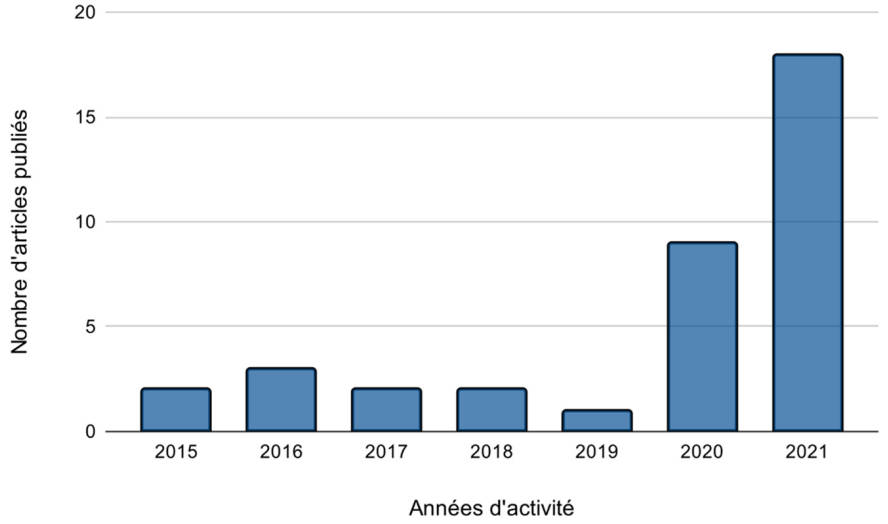


Figure 3. Nombre de publications du *JOHD* par an

29. Le *JOHD* a moins de dix ans d'activité, ce qui rend difficile l'obtention de statistiques sur une perspective à long terme. Cependant, nous pouvons mesurer la croissance de la revue en vérifiant combien d'articles ont été publiés au cours de ces sept années d'activité. Entre 2015

30. https://openhumanitiesdata.metajnl.com/collections/computational-humanities-research. L'appel à soumissions pour ce numéro thématique est désormais fermé.

31. *WMA Declaration of Helsinki - Ethical Principles for Medical Research Including Human Subjects*. URL : https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (en date du 16 août 2021).

et 2019, le *JOHD* a publié un à trois articles par an, en raison de son lancement récent, ce qui ne lui avait pas encore donné la possibilité de s'imposer dans le domaine des SHS. Toutefois, la revue a connu une croissance rapide, comme le montrent les 19 articles publiés en 2021 par rapport aux neuf articles publiés en 2020 (voir la figure 3). Nous précisons ici que ces chiffres se réfèrent à une étude qui porte sur la période 2015-août 2021 ; le nombre total de publications depuis 2015 jusqu'à fin 2021 est de 49 articles.

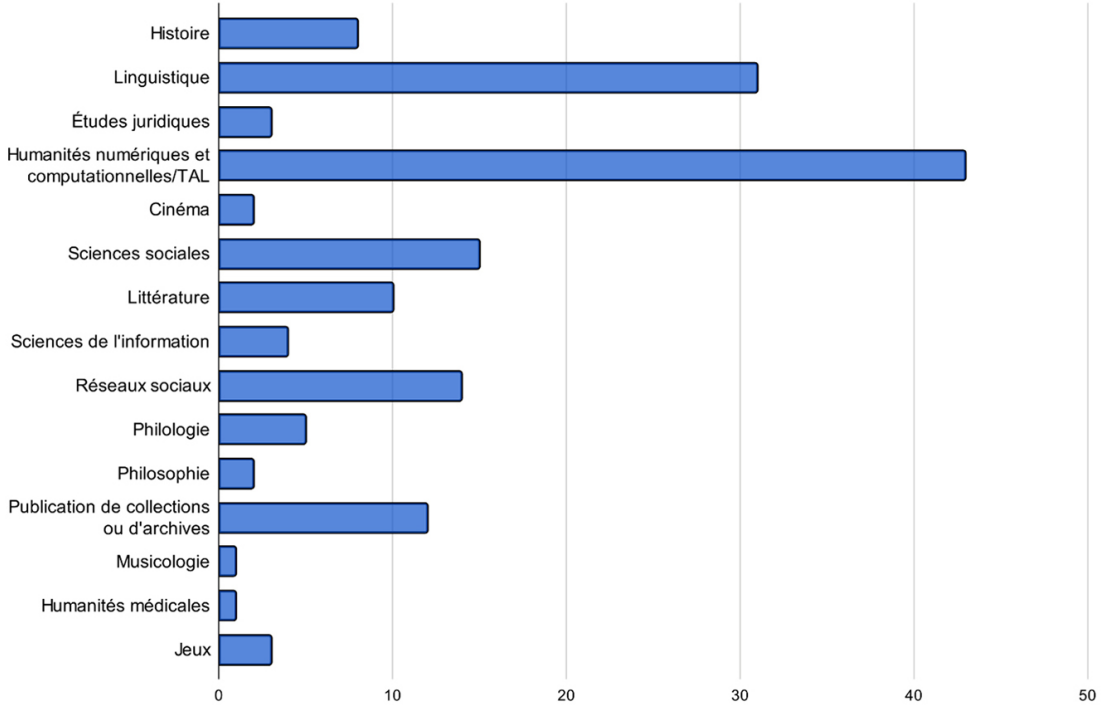


Figure 4. Les domaines des publications du *JOHD*

30. Les auteurs sont tenus de fournir des mots-clés lorsqu'ils soumettent un *data paper* ou un article de recherche. La revue ne fournit pas de liste de mots-clés, les auteurs peuvent donc choisir librement ceux qui leur semblent les plus représentatifs de leur article. Ils peuvent saisir jusqu'à cinq mots-clés destinés à indiquer les domaines, les sujets ou les techniques abordés dans l'article en question. Nous avons récupéré tous les mots-clés utilisés jusqu'à présent et les avons organisés en différents groupes en fonction de leur thème principal ou de leur sujet. De cette façon, nous avons pu définir les domaines qui sont les plus productifs en ce qui concerne la publication d'articles axés sur les données et repérer ceux dans lesquels ce type de publication doit encore se développer (figure 4). Jusqu'à présent, nous avons identifié 15 domaines de publication : histoire, linguistique, études juridiques, humanités numériques et computationnelles/TAL[32], cinéma, sciences sociales, littérature, sciences de l'information, réseaux sociaux, philologie, philosophie, publication de collections ou d'archives, musicologie, humanités médicales, jeux. Nous sommes conscients que ces libellés ne sont pas nécessairement comparables, car ils couvrent à la fois des disciplines et des méthodes. Cependant, l'analyse des mots-clés a également révélé que la plupart des articles sont multidisciplinaires : souvent, les mots-clés saisis par un même article couvrent plus d'un domaine. Les domaines les plus productifs sont sans aucun doute la linguistique et les humanités numériques

32. Traitement automatique des langues.

et computationnelles/TAL. La catégorie des études sociales couvre les articles qui traitent de COVID-19, sur le plan d'effets sociaux et culturels sur des communautés spécifiques (Hall-Lew *et al.*, 2021), d'impact sur les médias sociaux (Allés-Torrent *et al.*, 2021), de diffusion de fausses nouvelles et de leur détection (Knuutila *et al.*, 2021). Ces articles ont été publiés dans le déjà mentionné numéro thématique « *Humanities Data in the time of COVID-19* ». Les domaines les moins explorés sont la philosophie, la musicologie et les humanités médicales, qui sont toutes représentées par un seul mot-clé. Le dernier domaine qui est apparu est celui des jeux, grâce à un article récemment publié qui décrit la collecte et l'analyse d'un corpus de textes d'instructions pour des jeux.

Les *data papers* : des catalyseurs de la réutilisation des données ?

31. Cette section présente une étude que nous avons menée pour répondre aux questions suivantes : existe-t-il une corrélation entre la publication de *data papers* et la réutilisation des jeux de données ? Quelle est la relation entre un *data paper* et les données qu'il décrit ? Quelles sont les implications de la citation d'un *data paper* par rapport à la citation du jeu de données publié dans un entrepôt ouvert ?

32. Tout d'abord, nous avons tenté de recueillir des informations sur les citations de *data papers* et sur les citations de jeux de données, dans le but de déterminer s'il existe une corrélation entre les deux. Cependant, les informations sur les citations de jeux de données se sont avérées très peu fiables en raison des incohérences évidentes entre le décompte des citations fournies par les entrepôts de données[33] et celles fournies par les bases de données tierces telles que *Dimensions*[34]. Cela n'est pas surprenant étant donné que plusieurs études sur la citation de jeux de données et de logiciels ont souligné l'absence de normes de citation communes pour les données par rapport aux articles de recherche traditionnels (I) et (II) que les mentions informelles (c'est-à-dire notamment des URL intégrées dans le texte ou placées dans une note de bas de page et qui renvoient vers les données) sont au contraire beaucoup plus répandues (voir, entre autres, Park et Wolfram, 2019 ; Yoon *et al.*, 2019 ; Li et Yan, 2018 ; Park, You et Wolfram, 2018 ; Hwang *et al.*, 2017 ; Li, Greenberg et Lin, 2016 ; Pan *et al.*, 2015). De plus, de manière plus générale, les concepts de réutilisabilité et de réplicabilité des données en SHS suscitent encore du scepticisme au sein de la communauté scientifique (Peels et Bouter, 2018 ; de Rijcke et Penders, 2018), ce qui témoigne de la nécessité d'encourager la discussion autour des bonnes pratiques en matière de science ouverte en SHS (Knöchelmann, 2019). Pour toutes ces raisons, un article de recherche traditionnel peut faire l'objet de nombreuses citations, alors que ce n'est pas nécessairement le cas, pour l'instant, pour un *data paper* ou un jeu de données en SHS.

33. Une alternative valide à l'utilisation du nombre de citations pour connaître dans quelle mesure un *data*

33. En plus, seulement les entrepôts principaux, tels que Zenodo et Figshare, fournissent des informations sur le nombre de citations des jeux de données déposés.

34. https://www.dimensions.ai

paper peut favoriser la réutilisation des données est l'utilisation de mesures autres que les citations, comme les consultations et les téléchargements, et des indicateurs Altmetric[35], qui donnent un aperçu plus large et plus complexe de l'impact de la recherche (voir Peters *et al.*, 2016 ; Bornmann, 2014). Nous avons donc cherché une corrélation entre le nombre de visualisations et les indicateurs Altmetric sur les *data paper*s publiés, et le nombre de téléchargements des ensembles de données respectifs. Les téléchargements de jeux de données sont sans doute révélateurs d'un engagement direct avec les données elles-mêmes (par exemple, des tentatives de reproduction des résultats d'un article). Nous nous sommes donc concentrés sur tous les *data paper*s publiés dans le *JOHD* jusqu'au 1er juin 2021. Nous avons ensuite effectué un test de corrélation de Spearman (tableau 1) sur deux paires de mesures : l'une entre le nombre de visualisations d'un *data paper* et les téléchargements du jeu de données correspondant, l'autre entre les tweets sur un *data paper* et les téléchargements du jeu de données correspondant.

34. Avec un coefficient de corrélation de rang de Spearman positif, mais non significatif (α = 0,05) entre les deux, nous n'avons pas été en mesure de tirer de conclusions définitives sur la relation entre les visualisations des *data papers* et les téléchargements des jeux de données. En revanche, il existe une corrélation statistiquement significative, modérée à forte (r = 0,52) entre le nombre de tweets mentionnant les *data papers* et le nombre de téléchargements des ensembles de données respectifs. Afin d'apprécier pleinement cette corrélation, il serait utile de comparer ces chiffres avec la relation entre les tweets mentionnant directement un jeu de données et ses téléchargements. Cette comparaison n'a cependant pas été possible[36], puisque de nombreux entrepôts publics (dont Zenodo et Dataverse) ne sont pas ou plus abonnés à Altmetric et n'affichent donc aucun chiffre sur l'engagement des réseaux sociaux à l'égard de leur contenu. Dans ce cadre, la décision de Zenodo de ne pas renouveler son abonnement à Altmetric en 2020 a restreint notre projet d'étude, puisqu'au moment de cette analyse, pour 37 % des *data papers* du *JOHD*, Zenodo avait été choisi comme entrepôt.

	r de Spearman	*p*-value
Visualisations (*data paper*) **vs. Téléchargements** (jeu de données)	**0.36**	0.14
Twitter (*data paper*) **vs. Téléchargements** (jeu de données)	**0.52**	**0.03**

Tableau 1. Test de corrélation de Spearman

Résultats d'un test de corrélation de Spearman entre le nombre de consultations des *data papers* du *JOHD* et les téléchargements de leurs jeux de données respectifs et entre les tweets concernant les *data paper*s du *JOHD* et les téléchargements de leurs jeux de données.

35. Les indicateurs Altmetric (https://www.altmetric.com) sont des mesures alternatives aux mesures traditionnelles, basées sur les citations. Les Altmetrics sont tirées du Web, et incluent (entre autres) les mentions dans les réseaux sociaux (par exemple X), les discussions dans les blogs de recherche, et les citations sur Wikipédia. Ils mesurent l'engagement du public envers les travaux scientifiques et peuvent donc être considérés comme un critère supplémentaire pour évaluer l'impact de la recherche.

36. En date du 17 août 2021.

35. Toutefois, la corrélation trouvée entre les tweets et les téléchargements de données est encourageante et montre qu'il existe un lien entre les *data papers* et la réutilisation des données. Par ailleurs, il semble raisonnable de penser que la communauté académique (y compris les auteurs des jeux de données analysés) est plus susceptible de tweeter à propos d'un article bien structuré et évalué par les pairs qui décrit un jeu de données, qu'à propos de données brutes déposées par les auteurs et qui n'ont pas fait l'objet de validation par les pairs. De même, il semble raisonnable de penser que la communauté X soit plus encline à réutiliser un jeu de données lorsqu'elle tombe sur un tweet concernant un article qui le décrit. En effet, l'évaluation par les pairs d'un *data paper* est généralement considérée comme une bonne indication que l'ensemble de données en question est non seulement utile sur le plan scientifique, mais aussi structuré d'une manière qui le rend réutilisable par d'autres chercheurs et chercheuses. Il reste à approfondir si la corrélation observée entre les tweets et les téléchargements du jeu de données est plus révélatrice de l'utilité des médias sociaux dans la promotion des ensembles de données ou bien de l'utilité des *data papers* pour la réutilisation des données.

Conclusion

36. L'objectif de ce chapitre a été de faire le point sur la publication de *data papers* pour les SHS et de situer la revue *JOHD* dans ce contexte, comme jouant un rôle actif dans la promotion des valeurs de la science ouverte et du partage des données. Tout d'abord nous avons donné un aperçu des différentes définitions de *data paper* existantes, en essayant d'en trouver une qui puisse tenir compte de son contenu, de ses objectifs et des critères pour définir et évaluer ce type de publication. Nous avons ensuite présenté un résumé de l'histoire de la publication de *data papers* au fil du temps, qui a révélé une différence importante entre les sciences dures et les SHS. Par conséquent, nous avons essayé d'identifier la raison de cette disparité, en nous concentrant notamment sur la difficulté de trouver une définition du terme « donnée » qui pourrait être commune à toutes les disciplines en SHS, ce qui génère inévitablement une certaine méconnaissance de la valeur de l'ensemble de données en tant que résultat de la recherche elle-même, et sur la possibilité d'obtenir une reconnaissance académique pour le travail qui se cache derrière. Après avoir établi le statut actuel du *data paper* en SHS, nous avons présenté la revue *JOHD*, son histoire, son équipe éditoriale et le type d'articles qu'elle publie. Une attention particulière a été accordée à la description des activités internes et externes menées par les membres de l'équipe éditoriale, non seulement pour faire connaître la revue mais surtout pour l'inscrire dans le sillage des valeurs de la science ouverte et du partage de données. Nous avons ensuite présenté notre expérience dans la publication de *data papers* en SHS au moyen de données que nous avons pu recueillir au cours des dernières années. À ce propos, nous avons montré une croissance progressive en termes de nombre de *data papers* publiés par an, qui a connu un essor significatif en 2021. Nous avons également identifié les domaines les plus productifs en

termes de publication (et donc de soumissions) de *data papers* (voir, par exemple, les humanités numériques et le traitement automatique des langues) et ceux qui sont émergents (voir, par exemple, les études cinématographiques et les humanités médicales). Enfin, nous avons essayé de mesurer l'impact des *data papers* sur la réutilisation des données qu'ils décrivent, en trouvant une corrélation positive entre les consultations des *data papers* et les téléchargements des bases de données correspondantes, ce qui nous amène à attribuer au *data paper* un rôle actif pour rendre les données plus faciles à trouver et plus accessibles. Pour conclure, nous avons défini un modèle de science ouverte, transparente et reproductible qui voit dans la pyramide de l'accès ouvert sa concrétisation : le *data paper* n'est qu'une des voies que la communauté scientifique peut prendre pour améliorer sa recherche en vue des valeurs de science ouverte et de partage de données. L'utilisation combinée d'entrepôts de projet, d'entrepôts de données, de *data papers* et d'articles de recherche traditionnels, tous librement accessibles, peut maximiser l'impact de chacun de ces moyens de publication vers une recherche plus ouverte et transparente.

Remerciements

37. Merci à Nicolas Larrousse pour la relecture et ses conseils. À Anastasia Sakellariadi et Brian Hole pour leur contribution à la section « Le *Journal of Open Humanities Data* (*JOHD*) ». Et à Malithi Alahapperuma pour l'aide à la collecte de l'ensemble de données utilisées dans la section « Les *data papers* comme catalyseur de la réutilisation des données ».

Les données utilisées pour les sections « Publier des *data papers* pour les SHS : l'expérience du *JOHD* » et « Les data papers : des catalyseurs de la réutilisation des données ? » se trouvent dans Zenodo (DOI 10.5281/zenodo.7624854)

Les données utilisées pour la section « Les data papers aujourd'hui : la pyramide de l'accès ouvert » se trouvent dans Figshare (DOI 10.6084/m9.figshare.12333290.v2)

Écrire des *data papers* en SHS, exemples et partage d'expériences

Un *data paper* en SHS : pourquoi, pour qui, comment ?

Victor Gay

Introduction

1. Destinée aux chercheurs en sciences humaines et sociales souhaitant se lancer dans l'écriture d'un *data paper*, ce chapitre propose un retour d'expérience prenant appui sur la production récente d'un *data paper* (Gay, 2021) et aborde les enjeux auxquels un auteur de *data paper* est souvent confronté. Il se penche tout d'abord sur les raisons méthodologiques pour lesquelles l'écriture d'un *data paper* peut être utile en sciences humaines et sociales. Il aborde ensuite le public auquel s'adresse ce nouveau format éditorial. Enfin, il propose quelques éléments pratiques afin d'aider les chercheurs à écrire leur *data paper*. Tout au long de ce chapitre, je prendrai appui sur mon expérience issue de la rédaction récente d'un *data paper* : « Mapping the Third Republic. A Geographic Information System of France (1870-1940) » (Gay, 2021). Ce *data paper* décrit un système d'information géographique de la France de la Troisième République – la base TRF-GIS[1]. Cette base de données met à disposition nomenclatures et *shapefiles* annuels correspondant aux circonscriptions administratives de France métropolitaine de 1870 à 1940. Elle décrit les circonscriptions administratives générales (départements, arrondissements, cantons) ainsi que les circonscriptions militaires, judiciaires, pénitentiaires, électorales, académiques, ecclésiastiques et les inspections du travail. Elle met aussi à disposition des nomenclatures annuelles établissant une correspondance entre chaque commune contemporaine et les circonscriptions auxquelles elle appartenait[2].

Un *data paper* en SHS : pourquoi ?

2. Pourquoi écrire un *data paper* ? Alors que les sciences humaines et sociales (SHS) connaissent un tournant quantitatif depuis une dizaine d'années, la valeur scientifique de la production de données reste peu reconnue : les comités d'évaluations privilégient encore l'article de recherche traditionnel tandis que les pairs réutilisent volontiers les données créées par d'autres sans pour autant leur accorder de citation. Ce manque de reconnaissance semble pourtant incompatible avec le travail chronophage que demandent la documentation du processus de production des données ainsi que leur mise en conformité avec les principes FAIR – deux éléments nécessaires à la reproductibilité des travaux de recherche. Dans ce contexte, le *data paper* constitue

1. TRF-GIS signifie *Third Republic France Geographic Information System.*

2. Le *data paper* est librement disponible sur HAL à l'adresse suivante : https://hal.archives-ouvertes.fr/hal-02951461. Les données sont accessibles sur le Harvard Dataverse à l'adresse suivante : https://dataverse.harvard.edu/dataverse/TRF-GIS.

un outil qui peut permettre aux producteurs de données de faire reconnaître leur contribution scientifique en rendant leurs données facilement citables, mais aussi en améliorant la pertinence ainsi que le périmètre de la réutilisation de leurs données.

3. L'usage des données a pris une place prépondérante dans la recherche en sciences humaines et sociales au cours des dix dernières années, en partie grâce à la production sans précédent de statistiques portant sur les faits sociaux et leur disponibilité via les catalogues de données en ligne tels que Progedo-Adisp[3]. C'est par exemple le cas en sociologie, une discipline au cœur des SHS : une analyse des 400 articles parus entre 2000 et 2020 dans la *Revue française de sociologie* révèle une nette tendance vers le quantitatif, si bien que depuis une décennie, plus de la moitié des articles dans cette revue contient au moins une table ou un graphique présentant des statistiques (figure 1). C'est aussi le cas en histoire, comme en témoigne le récent numéro des *Annales. Histoire, sciences sociales* consacré à l'histoire quantitative (Karila-Cohen *et al.*, 2018), ainsi que le succès de l'ouvrage *Méthodes quantitatives pour l'historien* (Lemercier et Zalc, 2008), réédité en version anglaise il y a peu (Lemercier et Zalc, 2020). De même, ce tournant quantitatif concerne les milieux anglo-saxons, aussi bien en histoire sociale qu'en histoire économique, où elle est plus ancienne : alors que la proportion d'articles avec au moins une table ou un graphique présentant des statistiques restait stable autour de 90 % dans les principales revues en histoire économique entre 2005 et 2020 (Cioni *et al.*, 2021, 24), celle-ci est passée de 5 à 13 % dans l'*American History Review* sur la même période (Ruggles, 2021, 14)[4].

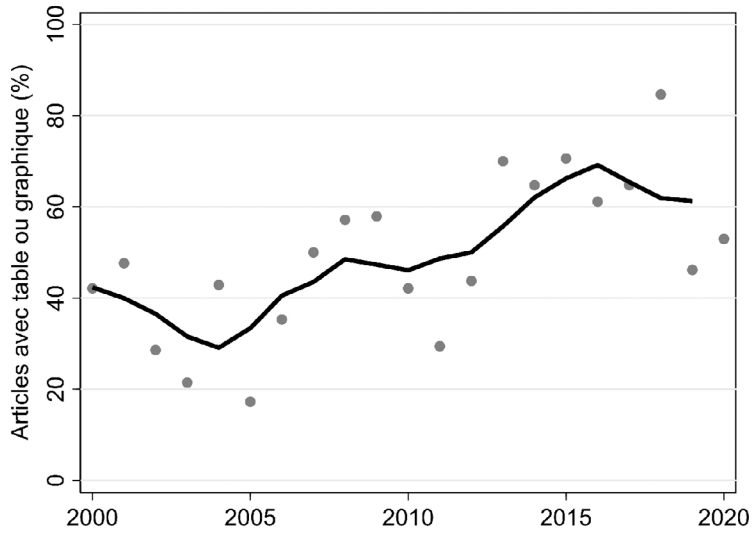


Figure 1. Pourcentage d'articles parus dans la *Revue française de sociologie* contenant au moins une table ou un graphique présentant des statistiques (2000-2020).

La courbe représente une moyenne mobile sur cinq ans. Les articles ont été récoltés par l'auteur via les portails Cairn.info pour les volumes 44 à 62 (2004-2021) et Persée pour les volumes 39 à 43 (1998-2002). Les éditoriaux, notes, critiques, débats, articles traduits, commentaires et les articles *in memoriam* n'ont pas été retenus.

3. Le catalogue PROGEDO-ADISP diffuse enquêtes et bases de données issues de la statistique publique française. Il est accessible à l'adresse suivante : http://www.progedo-adisp.fr.

4. Les journaux considérés par Cioni *et al.* (2021) sont l'*Economic History Review*, le *Journal of Economic History*, *Explorations in Economic History*, la *European Review of Economic History* et *Cliometrica*.

4. C'est dans ce contexte de renouveau quantitatif que la crise de la reproductibilité, partant de la psychologie et de la médecine, a rattrapé les sciences sociales, en commençant par l'économie (Maniadis, Tufano et List, 2017) : à cause des divers biais affectant le processus de publication – manipulation des valeurs p, faible puissance statistique, biais de confirmation des auteurs et relecteurs, etc. – une majorité des résultats de recherche publiés constitueraient en réalité des « faux positifs », autrement dit, une illusion statistique (Ioannidis, 2005). Plusieurs réponses ont été apportées pour pallier ce problème : les plans de gestion de données, les plans de pré-analyse, ou encore les méta-analyses (Maniadis et Tufano, 2017 ; Christensen et Miguel, 2018). Il semble cependant que la condition *sine qua non* de sortie de crise est la reproductibilité en tant que telle, c'est-à-dire la capacité à reproduire les résultats des travaux de recherche publiés. Tel n'est pas encore le cas : par exemple, Chang et Li (2022) montrent que seule la moitié d'un échantillon de 67 articles de macro-économie publiés dans des revues de renom sont reproductibles. La première étape qui découle de ce programme consiste donc en la mise en place des conditions pour la réutilisation (adéquate) des données de la recherche.

5. Comment y parvenir ? C'est ici qu'interviennent les principes FAIR, selon lesquels la reproductibilité requiert en premier lieu des données trouvables, accessibles, interopérables et réutilisables (Wilkinson *et al.*, 2016). Le respect de ces principes est depuis quelques années déjà au cœur de la politique nationale de la recherche française au travers du plan national pour la science ouverte ainsi que celle du financement de la recherche par l'ANR (CoSO 2019). Cependant, quel que soit le support des établissements et des infrastructures de recherche, le poids de la documentation et de la mise en conformité des données aux principes FAIR revient *in fine* en grande partie aux producteurs de données eux-mêmes, les chercheurs – qui sont par ailleurs déjà débordés par des tâches administratives qui ne cessent de s'accumuler (Ali et Rouch, 2013). Ce rôle est pourtant en contradiction avec les incitations auxquelles ils font face. En effet, la production de données n'est généralement pas valorisée par les comités d'évaluation, qui se fient encore bien plus aux publications traditionnelles dans des revues à comité de lecture (Gozlan, 2016). La production de données n'est pas plus valorisée par les pairs, qui en général ne citent pas les données qu'ils réutilisent. Par exemple, Robinson-García *et al.* (2015) montrent que seuls 18 % des données réutilisées par les articles en sciences sociales publiés entre mai et juin 2013 et disponibles dans *Web of Science* étaient citées explicitement dans leur bibliographie[5].

6. Face à ces défis, les *data paper*s offrent des perspectives intéressantes. En effet, en décrivant le processus de production des données dans un article publié dans une revue, ils offrent aux réutilisateurs de données un moyen simple pour citer les données, permettant par-là la reconnaissance scientifique du travail des producteurs

5. Une étude plus récente portant sur les pratiques de citation des données en biodiversité en 2019 constate que sur un échantillon aléatoire de 100 articles réutilisant de telles données, seuls 27 % les mentionnent explicitement dans leur bibliographie (Khan *et al.*, 2021).

de données par les comités d'évaluation, mais aussi par les pairs – au-delà, bien entendu, d'une utilisation à la fois plus adéquate grâce à la documentation et plus accessible grâce à la conformité aux principes FAIR[6]. De plus, dans la mesure où les *data paper*s passent au crible du comité de lecture, ils génèrent des incitations pour les producteurs de données eux-mêmes à parfaire leurs données ainsi que leur description, ce qui *in fine* ne peut que favoriser leur diffusion et le périmètre de leur réutilisation (Walters, 2020). Dans ce sens, le *data paper* peut contribuer à résoudre le problème classique du passager clandestin qui caractérise la production de biens publics purs – ici, la donnée documentée et FAIR.

7. Le *data paper* « Mapping the Third Republic » (Gay 2021) répond à ces problématiques dans un contexte de recherche particulier, celui de la Troisième République (1870-1940). En effet, cette période historique est caractérisée par une production sans précédent de statistiques (Desrosières, 2010 [1993]) par des administrations opérant à des niveaux d'agrégation hétérogènes, dans un contexte de profonds changements socio-économiques tels que la Seconde Révolution industrielle (1870-1914) ou la Première Guerre mondiale (1914-1918). Ainsi, grâce aux nouveaux moyens de numérisation des archives statistiques, la recherche en histoire quantitative sur cette période est en forte expansion. L'analyse et la visualisation de données historiques géolocalisées requièrent cependant un cadre de référence ou un système d'information géographique (SIG) commun. Celui-ci n'existant pas à ce jour pour la France de la Troisième République, chaque chercheur doit s'atteler individuellement à cette tâche. Cela implique une perte de temps considérable, des approximations à cause de la difficulté de l'entreprise, mais aussi un manque d'interopérabilité avec d'autres programmes de recherche et donc *in fine* un manque de reproductibilité. En effet, ces systèmes d'information sont rarement mis à disposition du public et plus rarement encore conformes aux principes FAIR[7]. La base de données TRF-GIS offre donc une solution à ces problèmes en proposant des données FAIR dont le processus de construction est documenté avec précision dans un *data paper*.

Un *data paper* en SHS : pour qui ?

8. Pour qui écrire un *data paper* ? Les utilisateurs potentiels des données décrites, bien sûr. Cependant, le lectorat d'un *data paper* est souvent bien plus large et interdisciplinaire qu'initialement envisagé par le producteur de données. En effet, les revues acceptant ce genre d'article restent rares en sciences humaines et sociales. À ma connaissance, il existe en 2021 deux revues dédiées aux *data paper*s – des *data journals* – dans ce domaine : le *Journal of Open Humanities Data* et le *Research Data*

6. Même si les effets des *data papers* sur la réutilisation des données n'ont pas encore atteint leur plein potentiel, en tout cas en sciences dures (Jiao et Darch, 2020).

7. Une exception toutefois : les *shapefiles* de cantons pour 1884 et 1925 publiés par le LARHRA (2011) et construits selon la méthode de vectorisation manuelle de cartes historiques géoréférencées. Voir Gay (2021, 14-15) pour une analyse critique de cette méthode.

Journal for the Humanities and Social Sciences – bien que certains *data journals* publient en sciences sociales mais aussi dans d'autres disciplines, comme *Data in Brief* ou *F1000 Research*[8]. De plus, quelques revues en sciences humaines et sociales acceptent des *data papers* en plus d'articles de recherche traditionnels, comme *Cybergeo* pour la géographie, *Historical Methods : A Journal of Quantitative and Interdisciplinary History* pour l'histoire ou *Frontiers in Sociology* pour la sociologie. Enfin, un certain nombre de revues offrent la possibilité de publier des *data papers* dans leur section pour articles courts sans qu'ils explicitent la distinction de ce genre d'article avec un article de recherche traditionnel. Pour le domaine de l'histoire, c'est par exemple le cas pour *Explorations in Economic History* ou encore *Histoire & mesure*.

9. L'auteur de *data paper* doit donc composer avec une offre réduite qui implique mécaniquement un lectorat plus large que son propre champ de recherche. Une telle portée demande cependant une adaptation de l'écriture à une audience non spécialiste : un travail d'exposition du contexte qui évite le jargon disciplinaire, une explicitation des présupposés ayant présidé à la collecte de données et aux catégorisations, et une clarification des usages potentiels des données au-delà de son propre champ de recherche. Ces considérations ont aussi des implications sur la forme des données : on peut bien entendu mettre à disposition ses données dans le format le plus courant dans sa discipline (spss, Stata ou sass) mais il est important de donner la possibilité à d'autres champs disciplinaires de s'en saisir en proposant aussi un format ouvert et universel comme le txt ou le csv. Le choix de l'entrepôt de données doit lui aussi être réfléchi : on peut ainsi se concentrer sur un entrepôt français comme Nakala ou Progedo-Adisp, ou bien viser un public plus international en utilisant le Harvard Dataverse, Figshare ou Zenodo – avec le risque de perdre en visibilité dans le paysage hexagonal, leur interface restant en anglais[9].

10. Enfin, il convient de clairement délimiter les contours du *data paper* en tant que descripteur de données et non comme procédant de l'analyse ou de la généalogie de projet, surtout si celui-ci est publié dans une revue n'ayant pas de section dédiée aux *data papers*. En effet, ce format n'est pas encore ancré dans les usages et nombre d'articles dont le but principal est de décrire une base de données comportent aussi des analyses des données elles-mêmes, ce qui contribue à brouiller les pistes[10].

11. Le *data paper* « Mapping the Third Republic » (Gay, 2021) propose nomenclatures et *shapefiles* annuels pour différentes administrations de la Troisième République (1870-1940). Dans la mesure où cette base de données

8. Une liste des revues publiant des *data papers* est proposée par Laurence Dedieu (Cirad) à l'adresse suivante : https://doi.org/10.18167/coopist/0057.

9. Il existe peu d'entrepôts spécifiquement dédiés aux sciences humaines et sociales, tels que Didómena (https://didomena.ehess.fr), l'entrepôt de données de l'EHESS dont l'usage est limité à ses membres, ou encore Nakala (https://nakala.fr/), l'entrepôt développé par la TGIR Huma-Num.

10. Par exemple, à la fin d'un article décrivant une nouvelle base de données sur les esclaves émancipés de la ville du Cap en 1834, Ekama *et al.* (2021, 9) analysent les déterminants des prix des esclaves par une régression linéaire.

permet de réaliser aussi bien de la cartographie de données historiques géolocalisées que d'appareiller des données issues de ces administrations pour réaliser des analyses statistiques, elle s'adresse à l'ensemble des sciences sociales avec une composante historique, et en premier lieu l'histoire économique et la démographie historique. C'est pourquoi mon choix s'est porté sur le journal *Historical Methods : A Journal of Quantitative and Interdisciplinary History*, une revue relativement reconnue, avec une diffusion internationale large, et dont le lectorat reste interdisciplinaire en sciences sociales avec une forte composante historique et une dominante quantitative. *Explorations in Economic History* aurait été une autre candidate, mais son lectorat reste trop restreint à l'économie historique anglo-saxonne. *A contrario*, le lectorat de *Journal of Open Humanities Data* ou de *Research Data Journal for the Humanities and Social Sciences* est relativement moins tourné vers la cartographie ou l'analyse économétrique.

12. En termes de format, la base de données *TRF-GIS* est disponible dans le format dominant du champ de l'histoire économique : le *Stata data format* (dta). Mais pour permettre à d'autres disciplines de se saisir des données, j'ai aussi rendu celles-ci disponibles en format txt, facilement importable dans n'importe quel logiciel statistique ou cartographique. Enfin, je me suis porté sur l'entrepôt Harvard Dataverse. Cet entrepôt propose un espace de stockage gratuit conséquent pour les chercheurs individuels, des facilités de dépôts, une gestion automatique des métadonnées, ainsi qu'une bonne ergonomie d'utilisation pour le téléchargement et la navigation. C'est aussi un entrepôt avec une très forte diffusion à l'international. Afin de ne pas manquer de toucher les communautés francophones, j'ai cependant pris soin de déposer les métadonnées de la base de données sur la plateforme data.gouv.fr[11].

13. Enfin, pour ce qui est de la forme du *data paper*, le passage du comité de lecture s'est passé sans encombre. Cependant, la remarque suivante de la part d'un relecteur anonyme montre bien le chemin qu'il reste à parcourir quant à ce type d'article : « La question est de déterminer si ce travail constitue un article. Je ne pense pas que ce soit le cas. Ainsi, je ne pense pas que la revue puisse accepter ce texte qui ne pose pas une problématique historique claire, mais qui n'est qu'une sorte de ligne directrice et la description d'un produit fini[12] ».

Un *data paper* en SHS : comment ?

14. Comment écrire un *data paper*, pour qu'il ne soit pas qu'un simple codebook, mais constitue une véritable clé d'accès pour la compréhension et la réutilisation des données décrites ? Une tendance actuelle est la

11. Les métadonnées sur data.gouv.fr sont disponibles à l'adresse suivante : https://www.data.gouv.fr/fr/datasets/systeme-dinformation-geographique-de-la-france-de-la-troisieme-republique-1870-1940.

12. « The question is to determine if this work constitutes an article. I do not think so. Thus, I do not think that the review could accept this text which does not ask a clear historical problem, but is only a kind of guideline and a description of an end product. »

génération automatique de *data paper*s à partir des métadonnées de la base de données (Schöpfel *et al.*, 2019). Bien que ce type de format reste bien moins chronophage qu'un *data paper* standard et permette une diffusion rapide via le moissonnage des catalogues de données, il ne semble pas permettre de réutilisation adéquate à cause du manque de description des données dont il fait preuve – cette méthode est donc à proscrire. La description des méthodes de construction et des choix de typologie doit être au cœur du *data paper*, chose que seul un humain est capable de réaliser.

15. Dans ce contexte, la première difficulté à laquelle est confronté un auteur de *data paper* lors de la rédaction est celle de la structuration du texte, car celle-ci ne correspond pas tout à fait à celle d'un article traditionnel. Un excellent modèle à suivre est celui proposé par le *data journal* en sciences dures *Scientific Data* du groupe *Nature*[13]. Bien que ce modèle s'adresse aux sciences dures, sa structure peut tout à fait s'adapter aux sciences humaines et sociales. La structure proposée commence par une première section (« Contexte et résumé ») qui décrit succinctement les données produites, leur contexte scientifique ainsi que leurs réutilisations potentielles. Une seconde section (« Méthodes ») décrit avec précision toutes les procédures utilisées dans le processus de production des données afin que celui-ci soit reproductible. Ensuite, une section « Fichiers de données » décrit chaque jeu de données associé avec le *data paper*. Cela comprend variables, noms de fichiers, localisation, ainsi que formats et poids numérique. Une quatrième section (« Validation technique ») présente les analyses ou procédures ayant permis de confirmer la validité des données décrites (en sciences humaines et sociales, il peut s'agir d'une confrontation avec différentes sources ou avec des données comparables). Une cinquième section (« Notes d'usage ») permet à l'auteur de décrire plus en détail les procédures de réutilisation des données ainsi que de développer quelques exemples. Enfin, une dernière section (« Disponibilité du code ») permet le cas échéant de décrire les procédures d'accès au code de reproduction de la base de données.

16. Une seconde difficulté à laquelle est confronté l'auteur de *data paper* concerne la mise à disposition des données. En effet, contrairement à l'article de recherche classique où les fichiers de données et de reproduction ne sont possiblement requis que lors de la phase de publication proprement dite, ceux-ci sont ici requis lors de la soumission de l'article à la revue avec des procédures d'accès qui doivent être explicitées dans le texte. Dans la mesure où c'est aussi l'intégrité des données qui est évaluée par les relecteurs, ceux-ci doivent impérativement y avoir accès dès l'étape de relecture. Les données doivent donc déjà avoir un identifiant provisoire à ce stade, ce qui permet de rattacher les données au *data paper* et de créer un écosystème « données-*data paper* ». Heureusement,

13. Les instructions de soumission à *Scientific Data* sont disponibles à l'adresse suivante : https://www.nature.com/sdata/publish/submission-guidelines. D'autres modèles existent, comme celui proposé par *Data in Brief*, disponible à l'adresse suivante : https://www.elsevier.com/journals/data-in-brief/2352-3409/guide-for-authors. Plus généralement, Kim (2020) propose un tour d'horizon des formats de *data papers* proposés par un certain nombre de *data journals*.

la plupart des entrepôts de données (par exemple le Harvard Dataverse ou Figshare) offrent la possibilité aux producteurs de données de générer un lien privé accessible (anonymement) par l'intermédiaire d'un code qu'il faudra fournir au comité de lecture pour l'évaluation par les pairs. Une fois toutes les modifications effectuées, les données seront accessibles au public via un identifiant pérenne, par exemple un DOI.

17. La rédaction du *data paper* « Mapping the Third Republic » (Gay, 2021) suit rigoureusement le modèle proposé par *Scientific Data* décrit plus haut. En effet, après une introduction présentant le contexte scientifique dans lequel la base de données s'insère, le corps de l'article consiste en une section « Méthode » qui explique en détail non seulement la méthodologie technique de construction ainsi que ses limites, mais aussi les éléments institutionnels qui sous-tendent les variations temporelles de chaque administration au cours de la Troisième République. J'y aborde par exemple les changements territoriaux impliqués par la perte et le retour de l'Alsace-Lorraine, la réforme des arrondissements de 1926, les réformes militaires de 1873-1874, ou encore les différentes lois de redécoupage électoral (les éléments précis sont étayés par des tableaux situés dans un appendice électronique d'une centaine de pages). S'ensuivent une description des quinze jeux de données de la base *TRF-GIS* (variables, espace de stockage, formats, licence), puis une exposition de la validation technique de l'ensemble, sous deux formes : une confrontation des sources secondaires utilisées avec un ensemble de 175 sources primaires (individuellement listées en appendice et disponibles en PDF dans l'entrepôt de données) et une validation de la méthode de construction des *shapefiles* par comparaison avec des données similaires (le *shapefile* des cantons de 1884 du LARHRA). L'article se termine par la description de l'emplacement du code et des données, ainsi qu'un exemple de réutilisation (la cartographie de l'abstention lors des élections législatives de 1914 au niveau des circonscriptions électorales).

Conclusion

18. La production d'un *data paper* demande un certain apprentissage dans la mesure où il s'agit d'un format nouveau en sciences humaines et sociales. Il est cependant en pleine expansion, et les chercheurs désireux de se lancer dans l'écriture d'un *data paper* ont à leur disposition de nombreux retours d'expérience à travers séminaires, journées spéciales, ou colloques dédiés (souvent mis en ligne, comme celui-ci). J'ai par exemple pu présenter le *data paper* décrivant la base *TRF-GIS* dans le cadre de la semaine DATA SHS 2020 organisée par la Plateforme universitaire de données de Toulouse (PUD-T) en décembre 2020, au webinaire dédié aux *data papers* organisé par le Groupe de travail inter-réseaux « Atelier Données » du CNRS en février 2021, au séminaire « Des sources aux SIG : des outils pour la cartographie dans les humanités numériques » en mai 2021, ou encore au

séminaire « Histoire et numérique » du Centre d'histoire de Science Po (CHSP) en mai 2021[14].

19. De nombreuses ressources sont aussi mises à disposition par différents acteurs institutionnels comme DoRANum et CoopIST (Cirad), qui proposent des dossiers sur le sujet, les URFIST ou l'INRAE, qui proposent des formations, ou encore les Plateformes Universitaires de Données (PUD) de PROGEDO, qui proposent des journées sur ce thème dans les MSHS lors de la semaine annuelle DATA SHS[15].

Pour consulter les données mobilisées dans le chapitre, voir :
Harvard Dataverse : https://dataverse.harvard.edu/dataverse/TRF-GIS
PROGEDO Adisp : http://www.progedo-adisp.fr/enquetes_donhist.php

14. La vidéo de ma présentation dans le cadre du séminaire « Des sources aux SIG » est disponible à l'adresse suivante : https://youtu.be/mBAlRdWR41k, de la minute 2 à 45.

15. Les ressources de DoRANum sur les *data papers* sont disponibles à l'adresse suivante : https://doranum.fr/data-paper-data-journal. Celles de CoopIST du Cirad sont disponibles à d'adresse suivante : https://doi.org/10.18167/coopist/0057. Un exemple de support de cours proposé par l'INRAE est disponible à l'adresse suivante : https://dx.doi.org/10.15454/1.4782473899888942E12. Voir aussi Reymonet (2017).

Data paper en humanités numériques : Adressbuch 1854

Mareike König, Gérald Kembellec et Evan Virevialle

Introduction

1. Dans la recherche historique, les bases de données sont souvent créées en accompagnement de projets de recherche. Leur création et leur indexation s'orientent la plupart du temps vers les thèmes centraux des projets. L'utilisation ultérieure des données par d'autres n'est pas envisagée, ou seulement tardivement. Pourtant, la plupart du temps, beaucoup de temps et d'argent ont été investis dans les projets, notamment par le biais de travaux manuels de saisie et d'indexation. Il en a été de même pour le projet « Annuaire des Allemands à Paris de 1854 », une base de données accessible en ligne depuis 2006. Des résultats essentiels du projet de recherche qui l'accompagnait avaient été publiés dans des publications classiques à l'époque. Ce qui manquait cependant, c'était une description de l'ensemble de données sur lequel repose la base de données. Le présent chapitre vise à combler cette lacune. Il décrit et documente, sous forme de texte et non de simples métadonnées, le jeu de données ainsi que les conditions et les réflexions fondamentales qui ont présidé à sa création, à sa révision et à sa republication de 2020/2021. Le *data paper* accompagne ainsi la curation et la *fairification*[1] des données, tout comme la nouvelle version du site Web et la cartographie qui l'accompagne. Il doit permettre aux historiennes et historiens de travailler avec les données, de les comprendre, de les enrichir, de les évaluer et de les réutiliser. Nous souhaitons en même temps faire connaître le projet et notre démarche de curation des données et de remodélisation du dispositif de consultation. Le destin du projet et de ses données est intéressant parce qu'il devrait être typique de nombreux projets de numérisation du début des années 2000. La procédure peut servir d'inspiration, sans toutefois vouloir ou pouvoir s'imposer comme procédure de *best practice*. Pour l'écriture du *data paper* à 6 mains, nous n'avons pas utilisé un modèle, mais nous avons mélangé des éléments d'articles en histoire classique avec des éléments qui s'orientent davantage vers la description de données et de métadonnées, afin de rendre le *data paper* et les données qu'il décrit accessibles à tous.

2. Au moment de la révolution de 1848, on estime que quelque 60 000 Allemands peuplaient les rues du Paris préhaussmannien et de ses banlieues (Grandjonc, 1974 ; Grandjonc, 1983 ; König, 2006 ; König, 2003). Par

1. L'institut historique allemand souhaite aujourd'hui libérer et diffuser les données ainsi que le code sous une licence ouverte pour permettre et encourager leur réutilisation dans la recherche. Voir les principes FAIR pour rendre les données de la recherche faciles à trouver, accessibles, interopérables et réutilisables par l'homme et la machine : https://www.ouvrirlascience.fr/fair-principles/.

Allemands, il faut comprendre immigrants temporaires ou permanents, intellectuels, artisans, ouvriers – qualifiés ou non – et domestiques de culture et de langue germanique. Ils étaient originaires des différentes principautés composant alors la Confédération germanique, de l'Autriche, de la Suisse ou d'autres régions où l'on parlait l'allemand. C'est le sort des habitants de cette « colonie germanophone », ainsi perçus par leurs contemporains, que le projet proposait d'étudier avec une approche d'histoire sociale (König, 2003). Les immigrés germanophones constituaient pendant longtemps au XIXe siècle, avant les Belges et les Italiens, le groupe d'immigrés le plus nombreux (Werner, 1995, p. 199-200 ; Noiriel, 1992). Ainsi, vers 1848, les Allemands représentaient plus de 35 % des étrangers à Paris. Compagnons et ouvriers suivaient une partie de leur formation à Paris et souhaitaient souvent en même temps échapper aux mesures politiques des gouvernements restaurateurs des États allemands pour vivre plus librement sous la monarchie de Juillet. À Paris, ils se sont associés à des militants politiques, des réfugiés, des artistes et des écrivains germanophones pour fonder le premier mouvement d'ouvriers allemands (Schieder, 1963).

3. La plupart des compagnons, apprentis et ouvriers allemands vivotaient dans les quartiers pauvres de Paris, souvent à la limite du minimum vital. D'autres, comme les artisans et les commerçants allemands, passaient par Paris pour y acquérir de nouveaux savoir-faire professionnels ou pour élargir leurs réseaux commerciaux. Nombre d'entre eux restaient, tout comme l'immigration bourgeoise et intellectuelle de l'époque, dans la capitale française pour profiter des possibilités économiques qu'elle offrait en tant que centre de l'art, de la mode et du savoir-vivre. Contrairement aux ouvriers non qualifiés et aux apprentis, ces riches artisans, commerçants et bourgeois allemands étaient en voie d'intégration. Ils apprenaient la langue du pays, épousaient des Françaises et demandaient parfois leur naturalisation (Dietrich et Varnier, 1987).

Rétrospective du projet et présentation de la source

4. Le projet de recherche sur l'immigration germanophone à Paris au XIXe siècle, lancé par l'Institut historique allemand de Paris en 2003, était accompagné par la mise en données d'un document unique à bien des égards, cofinancé à l'époque par la fondation Gerda Henkel : *l'Adressbuch der Deutschen in Paris für das Jahr 1854*[2] (Kronauge, 1854). Il ne reste aujourd'hui que quatre exemplaires de cet ouvrage, dont un conservé à la Bibliothèque historique de la Ville de Paris[3].

2. Traduction du titre en français : Annuaire des Allemands à Paris de 1854.

3. Selon l'OCLC l'ouvrage n'est accessible qu'à Paris (à la BnF et à la BHVP), Londres (*British Library*) et *Bonn Bibliothek der Friedrich Ebert Stiftung*. Voici l'url ark de la notice du document archivé sous la cote 703983 à la Bibliothèque historique de la ville de Paris. https://bibliotheques-specialisees.paris.fr/ark :/ 73873/pf0000884072

5. Cet annuaire compile sur 248 pages[4] un total de 4 772 adresses de particuliers et d'entreprises germaniques domiciliés à Paris et dans sa banlieue en 1854. On y trouve notamment les membres des classes moyennes et de la bourgeoisie, avec un éventail des professions et métiers extrêmement large : surtout des artisans comme des ébénistes, menuisiers, tailleurs, orfèvres et imprimeurs, mais également des entrepreneurs et des négociants, ainsi que des libraires, des banquiers, des professions libérales comme des médecins, écrivains, architectes, artistes et musiciens. En compilant les adresses, l'initiateur, un certain F.A. Kronauge, dont nous savons seulement qu'il était professeur de langue et qu'il habitait rue de Richelieu, entendait « offrir à nos compatriotes un moyen de se retrouver et de se faire connaître les uns des autres » (Kronauge, 1854). La source est d'une grande importance pour la recherche historique : on y trouve 4 772 particuliers des 12 245 immigrés allemands recensés officiellement en 1851[5]. Parmi eux, notons des personnes célèbres comme le poète Heinrich Heine, l'architecte de la gare du nord Jakob Ignaz Hittorff, le libraire Klincksieck et quelques membres de la famille de Rothschild. Mais encore l'importance de la source réside dans le fait qu'y sont répertoriées des personnes inconnues, qui représentent un échantillon de l'immigration allemande aisée. Ces noms peuvent constituer le point de départ pour une recherche sur des parcours migratoires individuels, en croisant les noms avec d'autres sources, comme les registres des mariages, des baptêmes et de l'état civil, des actes de notoriétés ou les actes de naturalisation et d'autres sources encore (König, 2003).

6. Ce projet, initié au début des années 2000, s'est achevé, dans sa version initiale, en 2006. Il s'agissait à l'époque d'un travail historique dans une mouvance innovante, pionnière même, si l'on considère que les humanités numériques, en tant que concept problématisé, datent de 2004 (Schreibman, Siemens et Unsworth, 2004). L'objectif était de créer une base de données et de la publier en ligne pour ainsi mettre à disposition des chercheuses et chercheurs, généalogistes et internautes, les noms et adresses des immigrés allemands compilés dans l'*Adressbuch*.

7. Les informations enregistrées manuellement par l'équipe de recherche dans les années 2000 ont été versées dans une base de données de type FileMaker. Il s'agissait de fiches mentionnant les noms, les prénoms, les adresses, les professions, les noms de rues et les arrondissements (d'avant 1860 et d'aujourd'hui[6]). Malheureusement, visualiser de manière interactive – avec des filtres – les adresses sur un plan de Paris de l'époque, représenter la

4. Une erreur de pagination l'amène en réalité à 242 pages. L'erreur se trouve dans les trois exemplaires qui ont été vérifiés au cours du projet : deux à Paris et à un Londres.

5. Le chiffre était probablement plus élevé que compté dans le recensement, voir Grandjonc, 1972. *L'Adressbuch* se vante d'être complet (*vollständig*), néanmoins une recherche rapide montre aussitôt les lacunes que les rajouts comme le fameux poète polonais Adam Mickiewics ou encore une librairie espagnole et une librairie polonaise.

6. Dans ce chapitre, nous mentionnerons les arrondissements d'avant 1860, en vigueur lors de la publication de la source principale.

répartition géographique des immigrés allemands était en 2003/2006 techniquement trop complexe. De plus, les données brutes n'étaient pas disponibles en *open data* sur le dispositif initial : à l'époque, on ne pensait pas à proposer le téléchargement des données ; le site Web et les données étaient placés sous copyright.

8. Environ 15 ans plus tard, le corpus est régulièrement consulté et fait l'objet de demandes d'informations, notamment de la part de généalogistes et des historiennes et historiens menant une recherche prosopographique. Le projet *Adressbuch 1854* a donc été relancé en 2020 et, depuis 2021, il bénéficie de l'aide de l'*Institut für Digital Humanities* de Cologne. Ce dernier fournit l'infrastructure d'hébergement du dispositif de consultation des données du projet. Les données brutes sont hébergées sur la plateforme Zenodo, le code du dispositif de consultation est déposé sur les comptes GitHub de deux institutions impliquées dans le projet.

Les finalités et enjeux du nouveau projet

9. La transformation de la base obéit à des objectifs et enjeux pluriels. Il s'agit avant tout de nettoyer, structurer et enrichir les données originelles pour permettre des usages plus larges en relation avec les besoins des historiens et historiennes, généalogistes et plus généralement des érudits et érudites. Ensuite, pour respecter les bonnes pratiques, nous désirons proposer un jeu de données qualifiées, accessible en *open data*. Les défis d'une modélisation en diachronie ont été nombreux : si les métiers et contextes socioprofessionnels ont évolué avec la révolution industrielle, il en va de même pour la terminologie associée. De plus, le paysage de la ville de Paris, sa toponymie et son découpage administratif ont subi de profondes mutations en 170 ans, notamment avec les travaux de Haussmann dans les années 1850/1860 (Gaillard, 1997 ; Pinon, 2012).

10. Ensuite, ce corpus mérite une remodélisation à visée humaniste pour correspondre aux sujets et méthodes des sciences humaines et sociales. Cette visée humaniste consiste à offrir une base de connaissances, en l'occurrence une base de données, dont la structure réponde à la fois à la vision historienne et à la possibilité de partage selon les standards FAIR des humanités numériques tout en respectant les règles de modélisation informatique. Pour questionner les données, il faut qu'elles soient finement articulées sous la forme d'un modèle qui valorise les problématiques historiennes. Chaque concept présenté dans la base est ainsi associé aux autres selon un point de vue historique. Les entités et associations telles que modélisées dans le paragraphe sous-titré « Penser les immigrés allemands dans un contexte socioprofessionnel et sur un territoire » seront à même d'éclairer le public cible selon ses propres critères. Pour faire face à la complexité du phénomène, appuyé conjointement sur plusieurs disciplines – tant sur le fond que sur la forme – le modèle structurant les données issues de l'*Adressbuch* a donc été pensé de manière transdisciplinaire. Une fine collaboration entre informaticiens, documentalistes et

historiens a permis d'en définir le fond, la forme et les méthodes d'accès. Il n'est pas possible ici de cloisonner le rôle de chacun : chaque aspect a été discuté pour répondre à la fois à une vérité sociohistorique, une réalité technique, aux besoins de consultation pluriels et enfin une méthode et de bonnes pratiques infodocumentaires : il s'agit là d'une réelle collaboration transdisciplinaire.

11. L'interface initiale du compagnon Web FileMaker ne correspond plus aux canons de consultation, d'ergonomie, d'accessibilité et de sémantique du Web. Il fallait donc repenser le dispositif de consultation en cohérence avec les nouveaux standards (design responsive, structuration sémantique des contenus…), les besoins de filtrages thématiques historiques, sociaux et cartographiques, mais aussi en cohérence avec les méthodes des humanités numériques.

Matériau de recherche et méthodes

Le modèle de données

Penser les immigrés allemands dans un contexte socioprofessionnel et sur un territoire

12. La base a été remodélisée dans une optique historienne avec l'outil de conception Mocodo puis implémentée sur l'application de gestion de bases phpMyAdmin. L'étape suivante consistait à créer les tables de données selon notre modèle. L'outil libre OpenRefine a permis de fragmenter le fichier tabulaire initial de données en plusieurs tables qui seraient ensuite exportées au format SQL : la base contient les tables courantes avec les données du projet (personnes, compagnies, professions ou rues…) et les tables associatives qui permettent la jonction entre les différentes tables courantes (ex. : la table qui permet la jonction entre les personnes et les compagnies : companies_persons). OpenRefine permet d'exporter les tables au format SQL pour ensuite les importer dans la base de données. En tout, ce sont 21 tables qui ont été créées et qui permettent d'interagir efficacement avec la base de données (voir figure 1). Une fois la base de données complète et stabilisée, elle fut transférée sur un serveur en ligne géré par l'*Institut für Digital Humanities* de Cologne.

Questions pragmatiques de transfert numérique

13. Comme spécifié précédemment, les données de la base initiale ont été exportées dans un format tabulaire, dans un fichier unique CSV encodé en UTF-8. Le CSV est un format qui ne nécessite pas de logiciels propriétaires pour être lu. Un simple éditeur de texte suffit pour lire un fichier CSV. Sa polyvalence est un avantage pour la diffusion et la réutilisation des données par les utilisateurs selon les principes FAIR. La norme d'encodage UTF-8 permet que le fichier soit rétrocompatible avec d'autres normes, comme la norme ASCII et il

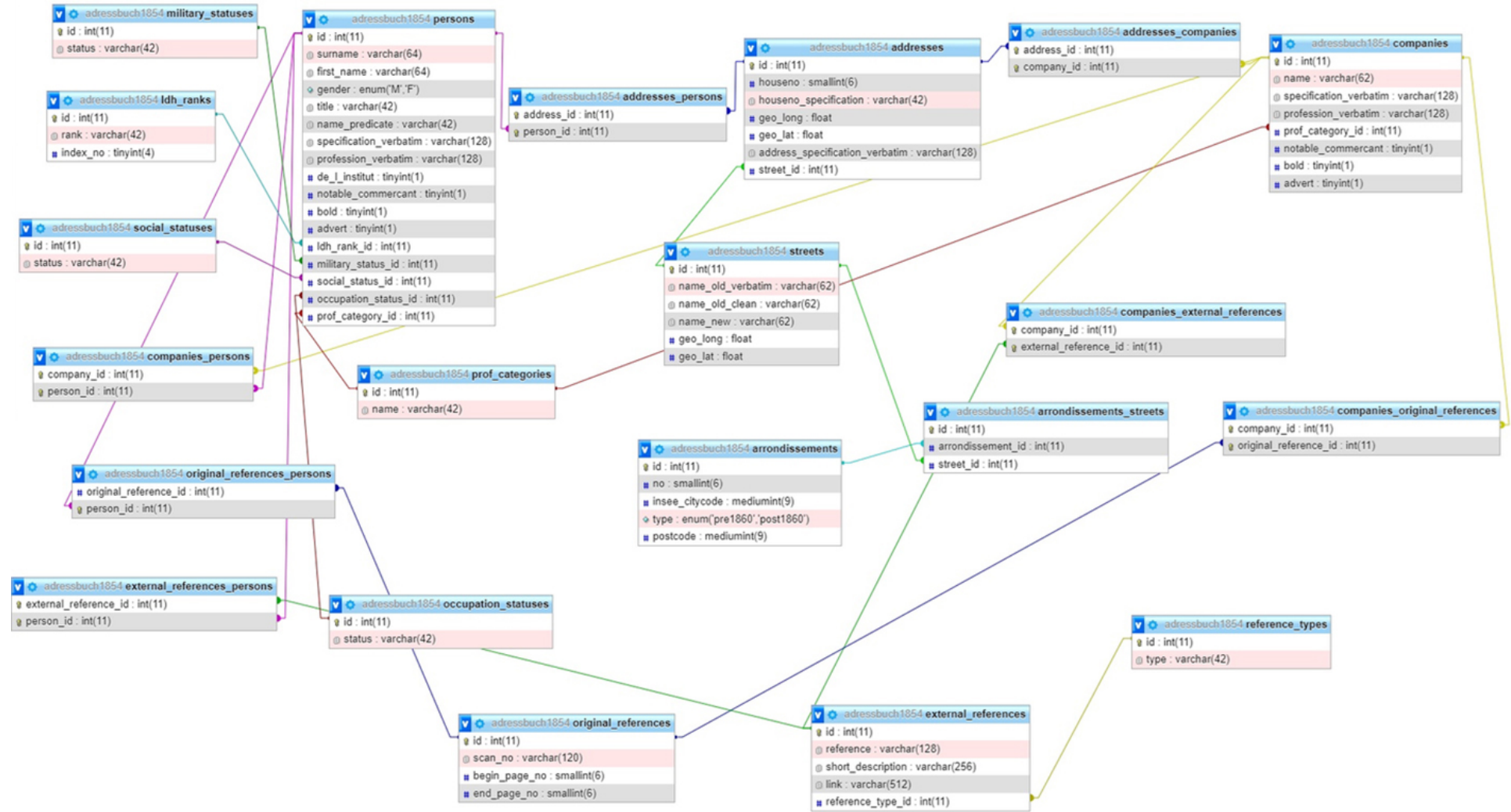


Figure 1. Le modèle de données.

Le modèle physique de données après une modélisation conceptuelle, licence *Creative Commons* BY-SA 4.0, DHI Paris.

permet de prendre l'ensemble du répertoire universel de caractères développé par l'ISO. Cela permet que tous les caractères spéciaux soient lus dans le fichier CSV. Elles ont été ensuite enrichies et réconciliées à l'aide d'OpenRefine avec des notices d'autorité et autres sources de données externes disponibles en ligne, comme les coordonnées GPS[7]. Le bruit (parenthèses, crochets…) a été éliminé des données avec le même logiciel au moyen d'expressions régulières du langage GREL[8]. Enfin, les données ont été segmentées pour correspondre au modèle présenté préalablement et ont été intégrées à une nouvelle base de données relationnelles.

Diplomatique numérique en histoire

14. Dans une optique historienne, la diplomatique se base sur le lien étroit qu'il peut y avoir entre les informations et l'archive : une base de données ne saurait se substituer dans ce contexte aux documents primaires qu'il convient de sourcer. Il est de plus indispensable d'associer les données et les métadonnées, mais aussi un *fac-similé* de qualité des pages de l'ouvrage. Il a été décidé de procéder à la numérisation des archives mobilisées pour offrir une alternative de qualité à la consultation physique.

15. La Bibliothèque historique de la Ville de Paris a mis à disposition de l'équipe le document original pour la numérisation. Le format d'encodage JPEG a été choisi pour les images numérisées pour sa compatibilité universelle et son ratio qualité-compression convaincant qui permet des impressions et visualisations de haute qualité (selon la résolution choisie). Les métadonnées de chaque numérisation ont été extraites pour permettre de proposer 2 tailles de vues différentes des pages numérisées : le format vignette pour l'affichage-écran et la haute définition pour les examens minutieux ou les tirages papier.

Documentation et classement de fac-similés

16. Les métadonnées techniques ont été extraites en ligne de commande avec un script Bash à l'aide de la commande Shell « file » pour obtenir les informations relatives aux images : la résolution, le format et la taille[9]. Ces métadonnées ont été associées au nom de l'image et à son répertoire de stockage selon les règles de nommage que nous définissons pour uniformiser et un plan de classement, puis enregistrées dans un fichier *Comma Separated Values* (CSV). Ces informations ont permis, grâce aux propriétés des images, de procéder à leur redimensionnement par lot. Pour l'affichage en vignette, nous avons choisi une résolution de 72 dpi et une taille de 400 x 800 px, soit la moitié de la taille des numérisations d'origine (voir figure 2). Le redimensionnement s'est également effectué en ligne de commande avec un script Bash et à l'aide de la bibliothèque Imagemagick[10] qui permet le traitement d'images en lots pour la modification de leurs résolutions et de leurs tailles. L'objectif

7. Par exemple : WikiData, BnF, GnD, ViaF, Geonames, etc.
8. Google Refine Expression Language, voir https://docs.openrefine.org/manual/grel.
9. Voir la documentation : https://linux.die.net/man/1/file.
10. Voir le site de l'application : https://imagemagick.org.

est de proposer un premier aperçu sans ralentir le chargement de la page de consultation avec une image volumineuse. Pour l'affichage en haute définition (HD), la résolution en 300 dpi a été retenue comme proposée par les règles de numérisation de la DFG[11]. C'est la résolution minimale pour proposer des visualisations et des numérisations de haute qualité compatible avec l'impression. Pour ne pas surcharger l'interface avec la taille et la résolution de la page numérisée, cette dernière s'ouvre dans un autre onglet lorsque l'on souhaite qu'elle s'affiche en plein écran[12].

Mise en texte de fac-similés

17. Les numérisations ont également été OCRisées[13] afin de proposer sur le site une vue affichant le texte d'une page (voir figure 2). L'OCRisation a été réalisée en ligne de commande à l'aide d'un autre script Bash et le package Tesseract[14]. Cette librairie permet d'obtenir le contenu de chaque numérisation au format texte avec un taux d'erreur acceptable sur les documents imprimés datant du XIX[e] siècle. Le jeu de données ainsi réalisé, déposé sur Zenodo en format ouvert (CSV, utf -8) et sous licence libre, se compose à la fois d'informations prosopographiques (métiers, adresses, grades militaires ou distinctions sont mentionnés) alignées géographiquement et d'une collection de numérisations de haute qualité.

Enrichir et aligner le jeu de données avec des autorités

18. Le travail de remodélisation s'est accompagné de l'alignement des anciennes données[15] avec le nouveau modèle au moyen du logiciel OpenRefine et de diverses autorités comme Wikidata ou encore des projets de recherche comme HISCO[16] qui recense et aligne en allemand et anglais les métiers anciens (Leeuwen, 2002).

19. OpenRefine permet un enrichissement des données avec une multitude de possibilités. Dans le cadre de notre projet, nous avons pu associer les coordonnées GPS contenues dans chaque fiche de rue depuis Wikidata pour les associer aux différentes adresses des personnes présentées dans la base de données. Le processus se fait de façon semi-automatique, c'est-à-dire qu'une partie est automatisée par OpenRefine et une partie est manuelle si le logiciel ne détecte pas de parfaite correspondance, ou détecte une ambiguïté. Des connaissances en histoire de la ville de Paris furent mobilisées pour s'assurer de la correspondance entre les rues de 1854 et celles d'aujourd'hui[17].

11. Voir le guide de bonnes pratiques « *DFG-PraxisregelnDigitalisierung* » : https://www.dfg.de/formulare/12_151.
12. Voir le dépôt principal du projet sur Zenodo : https://doi.org/10.5281/zenodo.7427439.
13. La reconnaissance optique de caractères, ou OCR, permet une traduction textuelle de caractères issus d'une image.
14. Voir la documentation présentée par Ubuntu : https://doc.ubuntu-fr.org/tesseract-ocr.
15. Les données du projet initial sont également partagées sous la forme d'un classeur, https://zenodo.org/record/4030877
16. Le projet, sa base et un outil permettant d'aligner jusqu'à 1 000 intitulés de métiers dans un même travail sont disponibles à l'URL https://historyofwork.iisg.nl
17. Un manuel très utile est : Jacques Hillaret, Dictionnaire historique des rues de Paris, Paris, 1963. Voir aussi le site « *La nomenclature officielle des voies parisiennes* » :

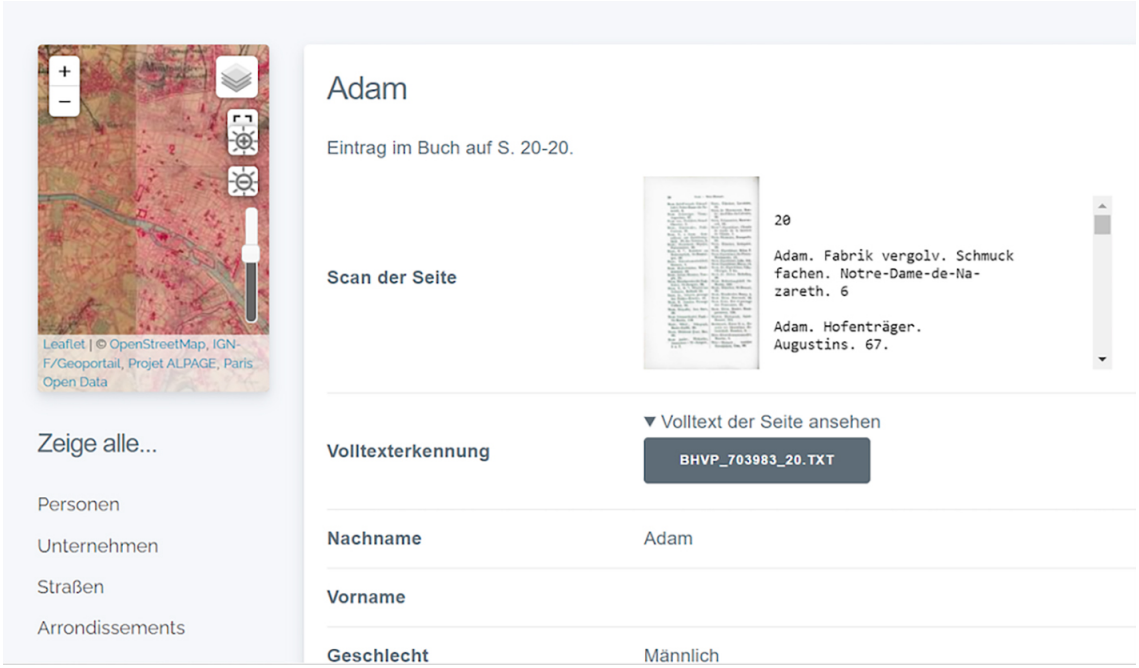


Figure 2. Rendu d'une page personnelle

Exemple de rendu d'une fiche d'une personne avec ses données personnelles issues de la base, le contenu textuel brut issu de l'OCR et la source primaire correspondante numérisée pour la gestion de la preuve. Licence Creative Commons BY-SA 4.0, DHI Paris.

Le dispositif de consultation

20. Nous avons évoqué précédemment que Web companion de FileMaker, l'outil pour créer la première interface, ne correspondait pas aux besoins des chercheurs d'aujourd'hui pour la consultation, le filtrage, la visualisation, l'accessibilité ou encore l'ergonomie du Web. Pour renouveler le projet, une interface suivant un modèle de conception *Model, View, Controller* (MVC) a été créée avec l'aide d'un *framework* Cake PHP.

21. Une nouveauté en rapport à l'interface de la première réalisation du projet est la cartographie. Il est possible de visualiser l'adresse des personnes et entreprises sur une carte de Paris grâce aux coordonnées géographiques calculées depuis l'association avec les données IGN. Pour afficher les latitude et longitude des entités de la base, une carte réalisée avec la bibliothèque de cartographie interactive Leaflet charge les coordonnées demandées au format JSON et positionne chaque individu ou société sur la carte.

22. Un des objectifs dans le cahier des charges de la suite du projet *Adressbuch* a été de proposer notre jeu de données ou des parties du jeu en téléchargement libre. Dans l'interface du dispositif de consultation, la base entière est proposée au téléchargement dans des formats libres : CSV, JSON, SQL ou XML. Les utilisatrices et utilisateurs peuvent également choisir de télécharger les données de plusieurs personnes par lot ou selon les résultats d'une recherche : des données filtrées par personne, ou groupe.

https://www.paris.fr/pages/les-voies-de-paris-denominations-et-numeros-d-immeubles-7550.

Pour une plus large diffusion des données du projet, la totalité est partagée dans les formats précédemment cités sur la plateforme de dépôt scientifique Zenodo sous une licence ouverte. Cela permet de rendre les données accessibles, découvrables, citables et réutilisables, grâce au *digital object identifier* (DOI) que la plateforme attribue aux jeux de données et à la licence CC-BY 4.0.

23. Dans une logique d'interopérabilité et une optique prosopographique à granularité fine, la fiche de chaque personne référencée dans la base est détectable par Zotero depuis le dispositif. Cette fiche peut donc être enregistrée par les chercheuses et chercheurs ou les généalogistes dans une base de connaissance de type Zotero avec des informations de type nom, prénom ou encore adresse. Cela peut être techniquement réalisé grâce à la norme NISO Z3988 (*Context Objects in Spans*) adaptée pour les besoins du projet à la plateforme CakePHP, ceci afin de correspondre au plus près aux recommandations d'ouverture et de description du FAIR.

Les résultats de recherche en méthodologie des humanités numériques

Sémiologie graphique et regard pluriel

24. Le renouveau du projet d'analyse de l'*Adressbuch* lancé en 2020 a permis de faire une étude diachronique de Paris, c'est-à-dire étudier la capitale française à deux époques différentes. Rappelons qu'en 1854, les arrondissements parisiens ne correspondent pas à l'actuel découpage qui date de la Loi du 16 juin 1859 pour la création des vingt arrondissements, avec le décret impérial du 31 octobre 1859 sur la dénomination des nouveaux arrondissements[18]. En outre, Paris a connu de profondes mutations à la suite des travaux de Haussmann de 1854-1870, avec une destruction en partie du vieux Paris au détriment des nouveaux grands axes et boulevards. Pour une parfaite compréhension de la problématique, les adresses des personnes ont été indexées selon les deux découpages et il est possible de choisir si l'on souhaite visualiser la carte selon une des deux modalités.

25. La carte numérique moderne OpenStreetMap[19] a été accompagnée d'un fond de carte historique provenant de l'état-major français entre 1820-1866. Cette carte est disponible dans le catalogue de cartes IGN issues du Géoportail du gouvernement[20]. Les données peuvent donc figurer et être visualisées au choix sur une carte moderne avec le découpage actuel et dans les contextes

18. Voir à ce propos cet autre document des archives de la ville de Paris : http://archives.paris.fr/_depot_ad75/_depot_arko/articles/47/correspondance-entre-les-arrondissements-anciens-et-nouveaux_doc.pdf

19. OpenStreetMap est une plateforme collaborative et ouverte qui possède une grande communauté de contributeurs en France et ne cesse d'évoluer. OpenStreetMap est le service de carte qui s'inscrit le mieux dans notre projet en suivant les principes d'open data et d'open source en opposition à Google Maps, par exemple, qui est propriétaire de ses données et payant pour certaines fonctionnalités.

20. Voir les données ouvertes IGN https://www.geoportail.gouv.fr/donnees/carte-de-letat-major-1820-1866.

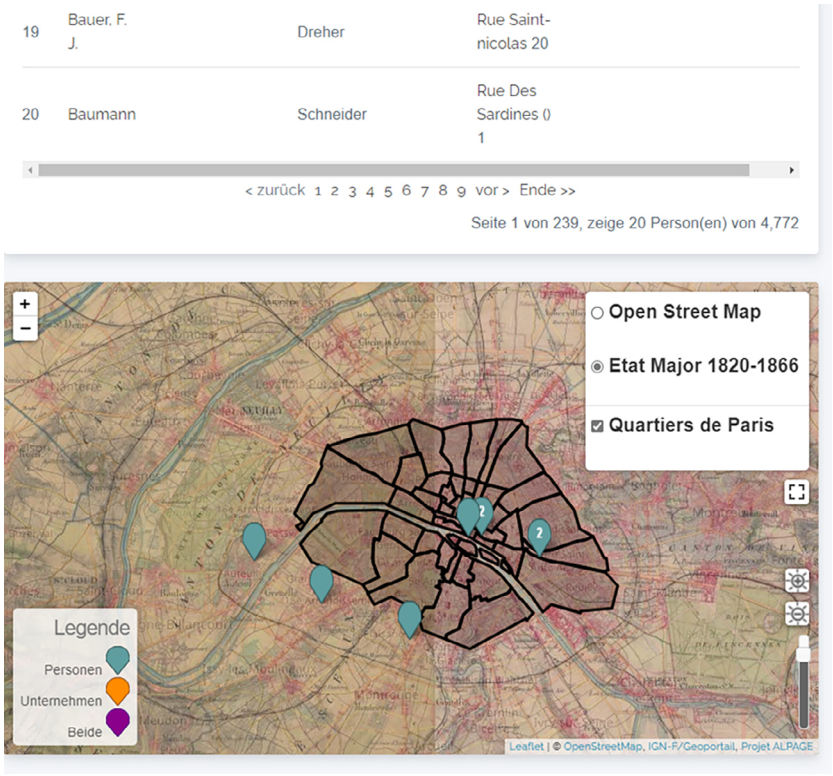


Figure 3. Carte diachronique des immigrés allemands à Paris en 1854

Exemple de filtrage de données utilisant Leaflet avec les données de l'*Adressbuch* géo référencées enrichies de celles du projet Alpage et une carte d'état-major IGN. Licence Creative Commons BY-SA 4.0, DHI Paris.

administratifs et topographiques historiques du Paris de 1854. Cette double vision en diachronie, avec la possibilité d'une surimpression en couches multiples, peut permettre d'interpréter l'impact historique potentiel du passage des Allemands à Paris sur la ville moderne et le contexte historique grâce à Opacity Controls[21], une extension de Leaflet. Mettre en place une cartographie multimodale avec un double découpage administratif est une nouveauté dans le cadre du projet. Cela se réfère aux principes d'expressivité graphique des données cartographiques tels qu'exprimés par Johanna Drucker (2011). Les données rassemblées, agencées, construites même, par le projet, ce que Drucker nomme « *capta* », peuvent être réalisées visuellement dans une relation de co-dépendance entre l'observatrice ou l'observateur et la carte : une analyse cartographique multiple est possible. Le choix d'une observation située en diachronie s'ouvre à l'utilisatrice ou l'utilisateur : soit une carte moderne, soit un fond de carte d'époque, soit également le découpage administratif effectif lors de l'observation initiale de la population des immigrés, soit celui qui est apparu quelques années plus tard et qui est toujours d'actualité, soit pas de découpage administratif du tout (voir figure 3).

Fiabilité des sources, reproductibilité des résultats

26. Dans un esprit d'ouverture et de partage des données, le code source partagé sur GitHub, en plus des simples

21. Voir https://github.com/lizardtechblog/Leaflet.OpacityControls.

données, permet de cloner la totalité du projet pour répliquer, s'approprier et modifier l'interface de consultation selon les besoins d'un nouveau projet : les focales spécifiques d'une autre discipline, la période étudiée ou encore les besoins de filtrage et/ou d'affichage[22]. Nous mettons à disposition les jeux de données primaires comme les numérisations de l'*Adressbuch* et la bibliographie qui nous a permis de mener à bien le projet. Puis, nous mettons également à disposition les jeux de données secondaires : la base de données avec les enrichissements issus des réconciliations avec les serveurs d'autorité ou encore les marqueurs cartographiques utilisés pour parfaire le dispositif de consultation. La véracité des données disponibles sur la plateforme en plusieurs formats est donc possible à vérifier et les potentielles erreurs peuvent être signalées. De plus, avec la publication des jeux de données, les analyses quantitatives induites présenteront des résultats qui pourront être reproduits, discutés, confirmés ou infirmés.

Conclusion

27. Mettre à disposition des chercheuses et chercheurs le contenu enrichi du livre *Adressbuch* leur permet d'explorer, de filtrer, d'analyser et de visualiser les données de façon plurielle et de les adapter à leurs propres besoins et questions de recherche. Ainsi, la base permet des observations statistiques et une représentation significative de la population allemande aisée au milieu du XIX^e^ siècle à Paris. Bien évidemment, sont explorées et visualisées ici seulement les données rassemblées par Kronauge, et non les données exhaustives sur l'immigration allemande à Paris en 1854. Néanmoins, considérant l'ampleur des données, leur représentativité et le fait que celles-ci portent en plus sur une époque sur laquelle la recherche historique repose souvent sur de pures spéculations, les résultats obtenus sont d'une grande importance. Ils permettent également de vérifier des thèses de recherche sur la migration et de corriger de petites erreurs, cette fois sur une base empirique.

28. La possibilité d'extraction et de téléchargement des données crée la condition préalable pour un *data mining* et une réutilisation individuelle. Pourtant, pour l'instant, les historiennes et historiens, en particulier, sont souvent très réticents à réutiliser les données provenant d'autres projets. Les principales réserves pour la réutilisation des données d'autres chercheurs sont la difficulté d'évaluer la véracité, la conformité aux sources et la qualité des données. Un *data paper* permet de rendre les données compréhensibles, tout comme les décisions prises pour les créer, enrichir et modéliser ainsi que les choix des formats et logiciels.

29. Nous avons rédigé ce data paper à trois, afin d'exprimer les différents points de vue disciplinaires sur le projet et ses données, et pour que le texte reste compréhensible à tous points de vue. Les valeurs d'une science ouverte sont la motivation fondamentale : un *data paper*, comme

22. Voir https://github.com/dhi-digital-humanities/Adressbuch1854.

discours d'escorte auctorial, rend les données compréhensibles en explicitant leurs structures, les sources associées, les conditions de collecte et de transformation. Nous souhaitons ainsi activement participer à un courant de promotion, par la réutilisation, des mises en données des travaux de recherche qui corresponde à une science durable et éthique. Enfin, nous souhaitions proposer et rendre possibles des scénarios d'utilisation divers des jeux produits, tant dans la sphère de la recherche que dans le monde de la culture : comme le dit Rufus Pollock, fondateur de l'Open Knowledge Foundation : « La meilleure chose à faire avec vos données sera pensée par quelqu'un d'autre »[23].

Logiciels utilisés

30. Mocodo : http://www.mocodo.net
31. OpenRefine : https://openrefine.org
32. Cake PHP: https://cakephp.org/
33. phpMyAdmin : https://phpmyadmin.net
34. Leaflet: Leaflet - a JavaScript library for interactive maps, https://leafletjs.com
35. Opacity Controls: https://github.com/lizardtechblog/Leaflet.OpacityControls
36. Shapefile Leaflet: https://github.com/calvinmetcalf/leaflet.shapefile
37. Commande « file » : commande shell Unix/Linux qui permet essentiellement de déterminer le type MIME d'un fichier. http://darwinsys.com/file/
38. ImageMagick : https://imagemagick.org/
39. Tesseract : https://doc.ubuntu-fr.org/tesseract-ocr
40. Projet Jupyter : https://jupyter.org/

Pour consulter les données mobilisées dans le chapitre, voir :

Référence primaire « Adressbuch » : F.A. Kronauge, *Adressbuch der Deutschen in Paris für das Jahr 1854. Vollständiges Adressverzeichnis aller in Paris und seinen Vorstädten wohnenden selbständigen Deutschen in alphabetischer Ordnung. Nebst Angaben der Sehenswürdigkeiten und Wohnungen der Gesandten*, Paris 1854. – Bibliothèques spécialisées et patrimoniales de la Ville de Paris, https://bibliotheques-specialisees.paris.fr/ark :/ 73873/pf0000884072

Jeux de données du projet « Adressbuch » : https://doi.org/10.5281/zenodo.7427439

Projet « ALPAGE » : AnaLyse diachronique de l'espace urbain PArisien : approche GEomatique – Alpage (huma-num.fr), https://alpage.huma-num.fr/

Géoportail, carte de l'état-major (1820-1866). Carte française en couleurs du XIX[e] siècle en couleurs, superposable aux cartes et données modernes. https://www.geoportail.gouv.fr/donnees/carte-de-letat-major-1820-1866

Correspondance des anciens (pré -1860) et nouveaux arrondissements de Paris : http://archives.paris.fr/_depot_ad75/_depot_arko/articles/47/correspon-

23. « *The best thing to do with your data will be thought of by someone else* », https://rufus-pollock.com/misc/.

dance-entre-les-arrondissements-anciens-et-nouveaux_doc.pdf

History of Work information system: https://iisg.amsterdam/en/data/data-websites/history-of-work

Adressbuch der Deutschen in Paris aus dem Jahr 1854, dispositif de consultation du projet « Adressbuch »: http://adressbuch1854.dhi-paris.fr/

Code source du projet « Adressbuch »: https://github.com/DH-Cologne/Adressbuch1854

Jupyter Notebooks utilisant les données du projet « Adressbuch »: https://zenodo.org/record/5512502#.Yi99zLjjJpR

https://github.com/dhi-digital-humanities/Adressbuch-Notebook

Utiliser un *data paper* en traitement automatique des langues : un exemple de classification automatique de mémoires et de thèses universitaires

Vincent Arnaud, Kevin Bouchard et Gilles-Philippe Morin

Introduction

Reproductibilité et réplicabilité

1. Afin d'être reconnue comme scientifique, une étude ne se doit-elle pas d'être reproductible ? Si la notion de reproductibilité est apparue dès le début des années 1990 (Barba, 2018 ; Peng et Hicks, 2021), cette dernière est au centre de débats récents dans de nombreux champs disciplinaires (Earp et Trafimow, 2015 ; Schmidt et Oh, 2016). À titre d'exemple, Baker (2016), rendant compte des résultats d'une brève enquête proposée par *Nature* auprès de 1 576 chercheurs, constate que « plus de 70 % des chercheurs ont essayé de reproduire les expériences d'un autre scientifique sans y parvenir, et plus de la moitié n'ont pas réussi à reproduire leurs propres expériences »[1]. Plus précisément, 52 % de ces chercheurs estimaient, en 2016, qu'il existait une crise majeure de la reproductibilité en science.

2. Il nous semble cependant important de distinguer deux termes voisins : reproductibilité et réplicabilité. En nous appuyant sur les propos de Peng *et al.* (2006), la reproduction d'une expérience vise à utiliser les données disponibles et à procéder à des analyses et interprétations indépendantes de l'étude initiale, quand la réplication consiste à collecter des données nouvelles et à effectuer des analyses conduisant à la confirmation ou l'infirmation des résultats antérieurs. Cependant, selon les domaines scientifiques, la signification de ces termes peut être précisée si l'expérimentateur, le plan d'analyse ou les routines d'analyse diffèrent ou non de l'étude initiale (Patil *et al.*, 2016 ; Plesser, 2018).

3. La reproductibilité et la réplicabilité d'une expérience impliquent que les chercheurs soient en mesure de partager (Laine *et al.*, 2007, 452) :

- leurs données : environ 45 % des chercheurs utilisent souvent ou la moitié du temps des données générées par d'autres, pour des méta-analyses par exemple (Science Staff, 2011). Le partage des données est une pratique scientifique éthique, de mieux en mieux admise par les éditeurs et valorisée par les bailleurs de fonds et la communauté scientifique. Cependant, même si le partage des données

1. « *More than 70 % of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.* »

s'est amélioré au cours de la dernière décennie, leur disponibilité et la volonté de les partager diffèrent encore beaucoup selon les disciplines (Tedersoo *et al.*, 2021). À cet égard, les humanités apparaissent comme l'un des domaines les plus partageurs.

- des éléments méthodologiques incluant, au minimum, le protocole expérimental et, le cas échéant, le code informatique utilisé pour produire les résultats. Comme le mentionne Berez-Kroeker *et al.* (2018, 4) « nous estimons qu'il n'est pas sain que les articles scientifiques soient étayés par des calculs qui ne peuvent être reproduits que par quelques employés d'un éditeur de logiciels commerciaux »[2]. Pour qu'une expérience puisse être reproduite ou répliquée, il semble donc essentiel, au-delà du protocole méthodologique, de rendre compte du code informatique utilisé pour effectuer les analyses. Cependant, même en disposant du code informatique nécessaire, le chercheur peut se voir contraint dans ses tentatives de reproduction ou de réplication par la puissance de calcul de son ordinateur. Par exemple, l'exécution d'un code informatique donné sur un puissant serveur informatique prendra nécessairement moins de temps que sur un ordinateur personnel[3].

4. Les travaux en linguistique ne font pas exception quant aux enjeux liés à la reproductibilité et la réplicabilité. En 2018, 41 chercheurs ont signé un article revendiquant que la reproductibilité s'applique aux « linguistes, notamment en ce qui concerne la promotion d'une culture de conservation à long terme et de citation des banques de données linguistiques »[4] (Berez-Kroeker *et al.*, 2018, 2). Pour leur part, Kobrock et Roettger (2022) ont cherché à évaluer, à partir d'un corpus de plus de 50 000 articles extraits de 98 revues en linguistique expérimentale, le taux de mention de la notion de réplication en recherchant, dans ces textes, la chaîne de caractères « replicat ». La distribution du taux de mention de cette chaîne de caractères varie beaucoup d'une revue à l'autre, allant de 0 à 12,82 %, le taux médian de mention de cette notion dans les 98 revues dépouillées n'étant que de 1,6 %. Ces auteurs révèlent également, à partir de l'analyse manuelle d'un sous-échantillon de 210 de ces quelque 50 000 articles, que seules huit études sont des réplications directes (3,8 %), c'est-à-dire des études qui parviennent aux mêmes conclusions scientifiques que l'étude initiale en utilisant la même méthodologie.

Articles exécutables, *notebooks* et reproductibilité

5. La dynamique mondiale de science ouverte entend répondre à la crise de la reproductibilité en science, non seulement en rendant disponibles les données, mais aussi en partageant les choix méthodologiques et les codes informatiques employés pour parvenir aux

2. « *Our view is that it is not healthy for scientific papers to be supported by computations that cannot be reproduced except by a few employees at a commercial software developer.* »

3. Les GAFAM (*Google, Apple, Facebook, Amazon et Microsoft*) disposent de puissances de calcul qui sont, à notre connaissance, difficilement accessibles au commun des chercheurs universitaires.

4. « *linguistic scientists, especially with regard to facilitating a culture of proper long-term care and citation of linguistic data sets* »

résultats. Trop souvent, même si les bases de données sont disponibles, le code informatique utilisé est absent ou inclus dans des scripts cryptiques qui ne sont pas construits de façon à ce que d'autres chercheurs puissent les relire et les réutiliser.

6. Un moyen de favoriser la disponibilité des données et la transparence des méthodologies employées est l'article exécutable (Lasser, 2020). Il s'agit d'un type de *data paper* qui permet de partager avec le lectorat chaque étape de l'analyse pour parvenir aux résultats obtenus, depuis les données jusqu'aux graphiques finaux. Par conséquent, ce document doit inclure les fonctionnalités nécessaires pour rendre le processus de recherche transparent et reproductible.

7. Plus précisément, un article exécutable combine du texte formaté, des liens, des références, un accès aux données, des figures, des résultats d'analyse, et éventuellement des vidéos et des éléments interactifs. Le code informatique utilisé pour créer les graphiques et réaliser les analyses doit aussi être affiché et interprété, permettant à l'utilisateur d'itérer l'analyse ou de la modifier. Enfin, un article exécutable doit être composé de composants entièrement libres et hébergés sur une plateforme accessible au plus grand nombre sous une licence gratuite pour être facile à partager et à réutiliser.

8. Les articles exécutables peuvent se présenter sous la forme de *notebook*s, parfois aussi appelés calepins électroniques. En informatique, les *notebook*s s'inscrivent dans le paradigme de la programmation lettrée (PL) (*Literate Programming*, Knuth, 1984)[5]. Tout comme les utilisateurs du langage de programmation *Python*, la communauté du logiciel *R* (R Core Team, 2022) a intégré la PL à son écosystème avec la création de *notebook*s (Gandrud, 2015 ; Xie, 2015). *R* est un langage de programmation et un logiciel libre largement utilisé en statistiques et en sciences des données. *RStudio* (RStudio Team, 2022) est, pour sa part, un environnement de développement facilitant l'utilisation du langage *R* notamment par une interface graphique.

9. En tant qu'objet technique, le *notebook* est écrit en *Markdown*[6], un langage de balisage très facile à apprendre qui permet de formater rapidement du texte de manière à le rendre facile à lire et à utiliser sur le Web (par exemple, le texte se trouvant entre deux astérisques sera interprété comme étant en italique). En plus du texte, le document est composé de sections (*chunk*) qui contiennent et interprètent du code informatique qui permet le chargement des données, leur nettoyage, leur sauvegarde, la production d'analyses statistiques, de figures ou de cartes (Stodden *et al.*, 2014). La figure 1 est une capture d'écran présentant une section incluant du texte formaté, une section de code informatique, par

5. La programmation lettrée vise à combiner le code source d'un programme informatique et la documentation explicative de ce code informatique dans le même document, les deux étant identifiés par des marqueurs spécifiques.

6. Pour un descriptif de la syntaxe de *Markdown*, le lectorat pourra se référer au texte de Gruber (2004) disponible en ligne à l'adresse suivante : https://daringfireball.net/projects/markdown/.

laquelle un lien pointant vers un document PDF est automatiquement extrait d'une page Web et une section dans laquelle est affiché le résultat de l'exécution du code informatique précédent. Le code affiché dans la seconde section est un code *R*, mais il aurait pu provenir d'autres langages de programmation.


En premier lieu, l'utilisation de différentes fonctions de la bibliothèque `rvest` permet de naviguer dans les nœuds et les attributs d'une page web pour en extraire une information spécifique. Dans ce cas, au sein de la page https://constellation.uqac.ca/3217/ présentant un mémoire de maîtrise en linguistique, nous cherchons à localiser le lien direct vers le fichier PDF contenant le texte dudit mémoire.

Hide

```
lien <- read_html("https://constellation.uqac.ca/3217/") %>%
                  html_nodes("[class='ep_document_link']") %>%
                  .[1] %>%
                  html_attr("href") %>% toString()
lien
```

```
[1] "http://constellation.uqac.ca/3217/1/RiverinCoutlxE9e_uqac_0862N_10109.pdf"
```

Figure 1. Capture d'écran extraite du *notebook* au format HTML accompagnant ce chapitre

10. En bref, un *notebook*, qu'il soit réalisé par l'entremise du logiciel *R* et de son environnement de développement *RStudio* ou de celle d'un autre langage de programmation, contiendra toujours des sections de code informatique, le résultat de l'interprétation de ce code et des informations en langage naturel qui rendent compte du processus de recherche.

Un exemple de *notebook* en linguistique

11. Dans un tel contexte où la reproduction ou la réplication d'une expérience nécessite non seulement le partage des données, mais aussi le partage des méthodologies et des techniques utilisées, cette contribution vise à rendre compte d'une analyse menée dans le cadre d'un projet de recherche en traitement automatique des langues avec la création d'un *notebook* par l'intermédiaire des logiciels *R* et *RStudio*.

12. Le jeu de données utilisé dans ce projet et le *notebook* contenant diverses précisions méthodologiques et le code informatique utilisé pour visualiser et analyser ces données sont disponibles de façon pérenne à l'adresse suivante : https://doi.org/10.34847/nkl.ad3b7717 dans l'entrepôt de données de recherche pour les Sciences humaines et sociales Nakala. Dans la suite de ce texte, sont présentés plus avant, le contexte, quelques points méthodologiques explicitant les choix effectués lors de la création du *notebook* et les résultats d'une expérience de classification automatique de textes en fonction du vocabulaire utilisé par les auteurs de ces textes.

Contexte et hypothèse de travail

13. La classification automatique de textes vise à assigner des documents textuels en langage naturel à une

catégorie donnée parmi un ensemble prédéfini de catégories (*multi-label classification*, Joachims, 1998). Elle est généralement approchée comme une tâche d'apprentissage dite supervisée, au cours de laquelle un algorithme de classification est entraîné à partir d'un ensemble d'exemples, puis utilisé pour classer des observations inédites (Sebastiani, 2002).

14. Le modèle classique utilisé pour une classification automatique de textes est basé sur la représentation en sac de mots (*bag of words* ou *BoW*). Cette représentation simplificatrice décrit un texte au moyen des unités lexicales qui le composent sans égard pour l'ordonnancement de ces unités ou les relations (autres que celles de cooccurrence) qu'elles entretiennent les unes avec les autres dans ce texte (Jurafsky et Martin, 2009, 57). Un texte est donc représenté par le vocabulaire qu'il contient sans tenir compte des relations grammaticales que les mots entretiennent entre eux.

15. Cette contribution est fondée sur ce type de représentation. Elle vise à tenter de classer automatiquement un échantillon de thèses de doctorat et de mémoires de maîtrise de l'université du Québec à Chicoutimi (UQAC) dans sept thématiques distinctes reflétant l'orientation générale des travaux (Administration et économie, Éducation, Lettres, Linguistique, Études régionales, Études théologiques, Travail social) en fonction du vocabulaire contenu dans ces textes. L'objectif final est d'évaluer la faisabilité d'un outil qui permettrait, à terme, à une bibliothèque universitaire de classer automatiquement un manuscrit dans différentes thématiques de recherche en fonction du vocabulaire contenu dans ce manuscrit.

Quelques points méthodologiques

Le corpus

16. Le texte brut de 1 030 thèses et mémoires de l'UQAC a été collecté automatiquement par moissonnage (*web scraping*) à partir des fichiers PDF disponibles au sein du dépôt institutionnel de cette université[7]. Le code informatique utilisé pour le moissonnage de ces ressources est présenté dans le *notebook*.

17. Cet échantillon ne regroupe que des thèses de doctorat (PhD) et des mémoires de maîtrise publiés en langue française sous forme de monographies. Ces travaux sont issus de quatre départements (Département des sciences de l'éducation, Département des sciences économiques et administratives, Département des sciences humaines et enfin Département des arts, des lettres et du langage, duquel ne sont extraits que les travaux en lettres et en linguistique) et de 18 diplômes (voir figures 2 et 3). Nous avons fait le choix d'exclure les travaux en arts, étant donné l'éclectisme de leurs formats.

7. Le dépôt institutionnel *Constellation* (https://constellation.uqac.ca) permet d'archiver, de diffuser et de valoriser la production intellectuelle de l'ensemble de la communauté de l'UQAC.

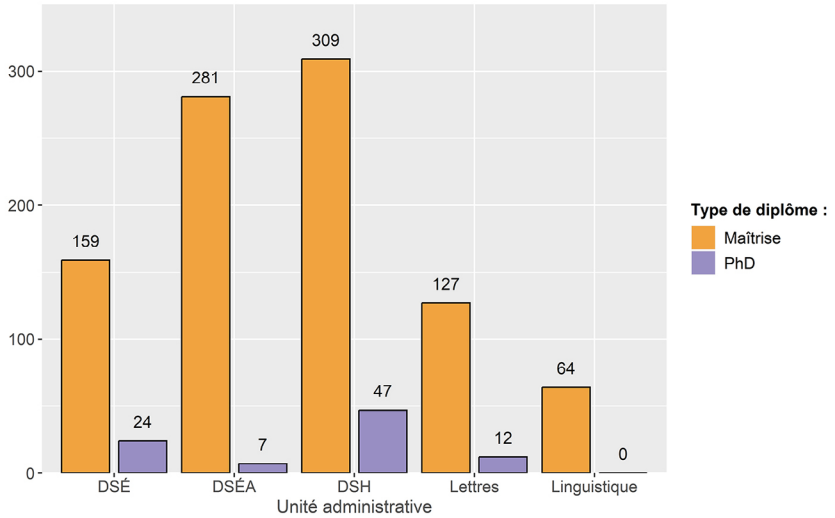


Figure 2. Distribution des 1 030 documents analysés par unité administrative et par type de diplôme (maîtrise ou PhD).

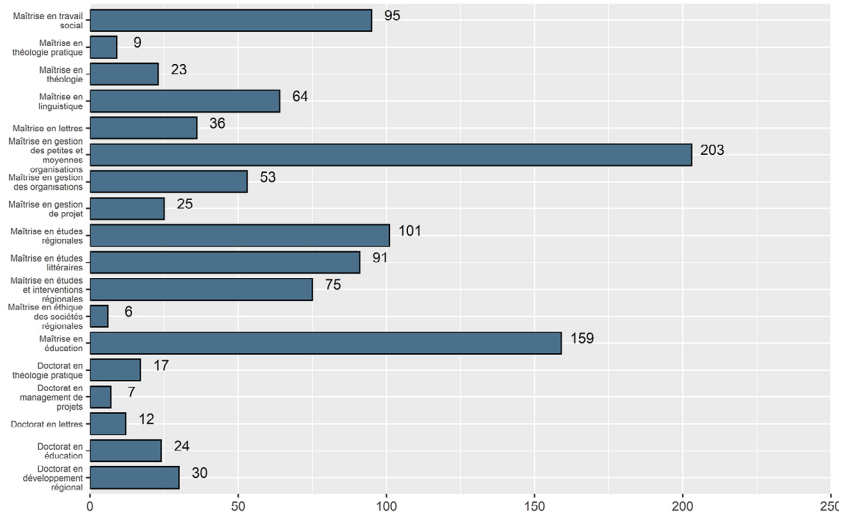


Figure 3. Distribution des 1 030 documents analysés par nom de diplôme.

18. En combinant, à la fois, la dénomination des diplômes et les découpages administratifs que constituent les départements, la catégorisation en sept thématiques de recherche présentée dans la figure 4 a été utilisée. Cette catégorisation ne tient pas compte de la distinction maîtrise/thèse de doctorat, étant donné le déséquilibre d'effectif entre ces deux types de diplômes, et ce, quelle que soit la thématique de recherche concernée (voir figure 1).

19. Les bibliothèques logicielles sous *R* (packages) *ggplot2* (Wickham, 2011) et *sjPlot* (Lüdecke, 2021) ont été utilisées pour construire les figures. Le code source est lui aussi indiqué dans le *notebook*.

Nettoyage du texte et analyse morphosyntaxique

20. En premier lieu, les textes bruts de chacune de ces 1 030 ressources textuelles ont été nettoyés automatiquement en utilisant différentes expressions régulières[8] qui ont permis de supprimer les pages liminaires, les annexes, les bibliographies, mais aussi de simplifier la mise en page (retours chariot intempestifs, tabulations, suites d'espaces induites par la présence de tableaux dans les fichiers PDF originaux…) (Welbers *et al.*, 2017).

8. Une expression régulière est une chaîne de caractères qui décrit, à l'aide d'une syntaxe précise, un ensemble d'autres chaînes de caractères ayant des propriétés communes. Une expression régulière permet, notamment, de rechercher des chaînes de caractères pour nettoyer du texte ou extraire des informations. Par exemple, l'expression régulière *replicat* permet de retrouver, dans un texte, toutes les formes commençant par la séquence de lettres *replicat* comme *replication*, *replicate*, *replicative*…

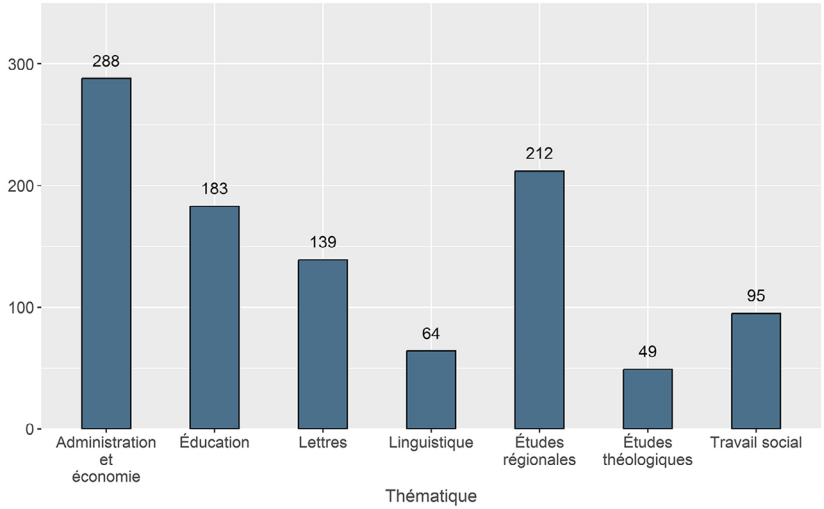


Figure 4. Distribution des 1 030 documents analysés par thématique de recherche

21. Une fois les documents nettoyés, l'étape suivante a consisté à les transformer en une représentation adaptée à l'algorithme de classification. Nous avons, pour ce faire, procédé à un étiquetage morphosyntaxique de chaque mot de chaque texte en utilisant le parseur morphosyntaxique[9] *UDPipe* (Straka et Straková, 2017), implémenté dans la bibliothèque logicielle *udpipe* sous *R* (Wijffels, 2022). Un exemple de code permettant l'analyse morphosyntaxique d'un texte est disponible dans le *notebook*.

22. L'étiquetage morphosyntaxique des mots composant chaque texte a consisté à leur attribuer un lemme[10], une catégorie grammaticale (*part of speech*) (verbe, nom commun, nom propre, adjectif, adverbe…) et un ensemble d'informations morphosyntaxiques (le genre et le nombre des noms ou des adjectifs, la personne d'un pronom, le temps d'un verbe…). Le jeu d'étiquettes est celui du projet *Universal Dependencies*[11] qui comprend notamment 17 catégories grammaticales possibles[12].

23. Pour être efficace, un analyseur morphosyntaxique automatique doit s'appuyer sur un modèle linguistique correspondant à la langue et à l'état de langue en usage dans les textes à analyser. Par exemple, il aurait été totalement inefficace d'utiliser le modèle linguistique UD-Old-French-SRCMF[13] qui repose sur l'analyse morphosyntaxique de plus de 18 000 phrases issues de dix textes d'ancien français du IX^e^ au XIII^e^ siècle. Dans le cadre de cette contribution, le modèle linguistique

9. Un parseur (ou analyseur) morphosyntaxique est un outil informatique permettant de découper une phrase en ses différents groupes syntaxiques, mais aussi d'attribuer des étiquettes morphosyntaxiques à chaque unité lexicale composant cette phrase (par exemple, dans le cas d'un nom, seront automatiquement indiqués son lemme, sa catégorie grammaticale, son genre et son nombre).

10. n lemme peut être défini comme la forme canonique d'une unité lexicale variable telle que mentionnée dans une entrée de dictionnaire. Il s'agira par exemple, pour un verbe, de sa forme à l'infinitif présent, et pour un adjectif de sa forme au masculin singulier.

11. Le projet *Universal Dependencies* (https://universaldependencies.org) vise de fournir un inventaire universel de catégories grammaticales et d'annotations morphosyntaxiques facilitant l'annotation de constructions similaires d'une langue à l'autre, tout en permettant des extensions spécifiques à une langue si nécessaire.

12. https://universaldependencies.org/u/pos/index.html

13. Conversion d'une partie du *Syntactic Reference Corpus of Medieval French* pour le projet *Universal Dependencies* : https://github.com/UniversalDependencies/UD_Old_French-SRCMF

UD-French-GSD a été utilisé[14]. Il repose sur l'analyse morphosyntaxique de plus de 16 000 phrases en français moderne couvrant plusieurs types de textes (nouvelles de *Google*, blogues, pages *Wikipédia*, commentaires d'utilisateurs…).

24. En utilisant ce modèle linguistique, chaque mot de chaque texte s'est donc vu attribuer automatiquement une catégorie grammaticale et un lemme. Pour chacun de ces lemmes, nous aurions pu calculer sa fréquence, c'est-à-dire le nombre d'occurrences de ce lemme dans chaque texte. Cependant, les mots les plus fréquents dans un document risquent aussi de l'être dans tous les textes de l'échantillon, c'est, par exemple, le cas des déterminants. Pour pallier cette difficulté, la valeur TF-IDF de chaque lemme (*Term Frequency-Inverse Document Frequency*) a été calculée. Un lemme, avec une valeur de TF-IDF élevée, est, à la fois, important dans un document donné, tout en apparaissant peu dans les autres documents de l'échantillon. Cette mesure statistique permet donc d'évaluer l'importance d'un mot en comparant le nombre de fois où il apparaît dans un document avec le nombre de documents dans lesquels il apparaît (Ramos, 2003 ; Welbers *et al.*, 2017).

25. On aboutit ensuite à une représentation de l'échantillon sous la forme d'une matrice documents/termes (*document/term matrix*) dans laquelle les documents sont en lignes, les lemmes en colonnes et les cellules indiquent la valeur TF-IDF de chaque lemme dans chaque mémoire ou thèse. Cette matrice de grande taille (sa taille correspond au nombre de documents × nombre d'unités lexicales retenues) est dite creuse (*sparse matrix*), car beaucoup de ses cellules contiennent une valeur TF-IDF nulle. Effectivement, les mots ne sont pas tous présents dans chaque texte. Il est alors possible de réduire la taille de cette matrice creuse en ne conservant que les lemmes étiquetés comme des noms communs et des adjectifs qui sont présents dans plus de 1 % des documents de l'échantillon. En utilisant cette technique de sous-échantillonnage, nous n'avons conservé que la valeur de TF-IDF de 15 928 lemmes présents dans les 1 030 documents colligés, ce qui demeure néanmoins un espace de représentation de très haute dimension.

26. Cette matrice creuse réduite a alors été divisée en deux jeux de données. L'un regroupant 70 % des thèses et mémoires de l'échantillon (n =720), destiné à entraîner l'algorithme de classification pour créer un modèle d'apprentissage et l'autre regroupant les 30 % restants (n =310), destiné à tester ce modèle d'apprentissage. Nous avons veillé à ce que les différentes décennies de publication, les thématiques et les départements d'attache soient présents dans des proportions voisines dans les échantillons d'entraînement et de test, afin de tester l'algorithme de classification sur des données voisines de celles qui ont servi à son entraînement.

27. Les matrices documents/termes des échantillons d'entraînement et de test sont également disponibles en

14. Conversion du corpus *Google Stanford Dependencies* pour le projet *Universal Dependencies* : https://github.com/UniversalDependencies/UD_French-GSD.

ligne[15]. Conjuguées à l'utilisation du *notebook* proposé, elles permettent de reproduire ou de répliquer le processus d'analyse.

Un modèle de classification automatique

28. L'utilisation d'un modèle de classification supervisée vise à estimer le degré d'adéquation de la catégorisation en sept thématiques aux lemmes des noms et adjectifs utilisés dans les travaux. Dit autrement, l'objectif est de classer automatiquement un texte dans une thématique définie *a priori* en fonction des lemmes des noms et des adjectifs utilisés par les auteurs de ce texte.

29. Le modèle de classification supervisée choisi est un modèle *XGBoost* (*eXtreme Gradient Boosting*[16], Chen et Guestrin, 2016). Il s'agit d'un algorithme disponible dans différents langages de programmation, Python et R notamment. Les atouts de *XGBoost* sont notamment liés à la possibilité d'ajuster de nombreux hyperparamètres pour un réglage fin du modèle d'apprentissage. Ce réglage fin permet d'augmenter les performances de l'algorithme lors de la tâche de classification qui lui a été assignée. Rhys (2020, 176) propose la définition suivante d'un hyperparamètre : « une variable ou une option qui contrôle la façon dont un modèle fait des prédictions, mais qui n'est pas estimée à partir des données »[17].

30. Il est important pour obtenir de bonnes performances de classification d'utiliser des valeurs optimales pour chaque hyperparamètre. Différentes méthodes mathématiques destinées à obtenir ces valeurs optimales parmi les valeurs candidates pour chacun des hyperparamètres peuvent être utilisées. Il serait, par exemple, possible de considérer cette optimisation comme un problème de recherche dans une grille. Par exemple, dans cette expérience, nous avons choisi d'optimiser la valeur de six hyperparamètres. Imaginons que nous établissons que chaque hyperparamètre peut prendre 5 valeurs, le modèle devra être entraîné 15 625 fois (5^6) pour obtenir une combinaison d'hyperparamètres pour laquelle les performances du modèle sont les meilleures, mais quelle garantie l'expérimentateur peut-il avoir que cette seule combinaison est optimale ? Les principales limitations d'une telle technique sont liées au temps de calcul, mais aussi à la granularité des valeurs candidates de chaque hyperparamètre.

31. Dans cette expérience, les valeurs optimales de ces six hyperparamètres ont été obtenues en utilisant une validation croisée à dix blocs[18] (*10-fold cross-validation*) jumelée à une optimisation dite bayésienne (Xia *et al.*, 2017)

15. https://doi.org/10.34847/nkl.ad3b77i7
16. https://xgboost.ai/
17. « *a variable or option that controls how a model makes predictions but is not estimated from the data.* »
18. Le principe général d'une validation croisée à dix blocs est le suivant : l'échantillon original est divisé en dix « blocs » (sous-échantillons), puis on sélectionne neuf de ces dix sous-échantillons comme échantillons d'apprentissage et le 10e comme échantillon de test. Après apprentissage, un score de performance est calculé avec l'échantillon de test. Puis, l'opération est répétée neuf fois supplémentaires afin que chacun des sous-échantillons joue le rôle d'échantillon de test. À l'issue de cette procédure de validation croisée, 10 scores de performance, un par bloc, sont disponibles, et il est possible de calculer une performance moyenne.

dont le code informatique est indiqué dans le *notebook* et qui peut être réutilisée dans le cadre d'autres expériences de classification automatique à l'aide de l'algorithme *XGBoost*. Les plages de valeurs des six hyperparamètres traités et leurs valeurs optimisées[19] sont aussi indiquées dans le *notebook* ce qui évite au lectorat souhaitant reproduire cette expérience de devoir recalculer les valeurs optimales de chacun de ces hyperparamètres.

32. Une fois les hyperparamètres optimaux définis, le modèle final a été entraîné à partir de l'échantillon d'entraînement représentant 70 % du volume total des ressources textuelles analysées. Le code informatique permettant l'entraînement de l'algorithme est détaillé dans le *notebook* disponible en ligne.

Résultats préliminaires

33. La matrice de confusion présentée dans la figure 5 et dont le code source est aussi disponible dans le *notebook* rend compte de la qualité de la tâche de classification automatique appliquée à l'échantillon de test. Dans cette matrice de confusion, chaque ligne correspond à la thématique de recherche à laquelle chaque document a été associé (cf. thématique *a priori*) et chaque colonne correspond à la thématique de recherche prédite (cf. thématique prédite) par l'algorithme de classification en fonction des lemmes des noms communs et des adjectifs présents dans chaque texte. Par exemple, parmi les 64 documents étiquetés comme études régionales, 52 ont été classés comme tels par l'algorithme de classification (soit 81,25 %). Sur la base du vocabulaire, l'algorithme prédit que les 12 documents restants appartiennent à cinq des six autres thématiques de recherche.

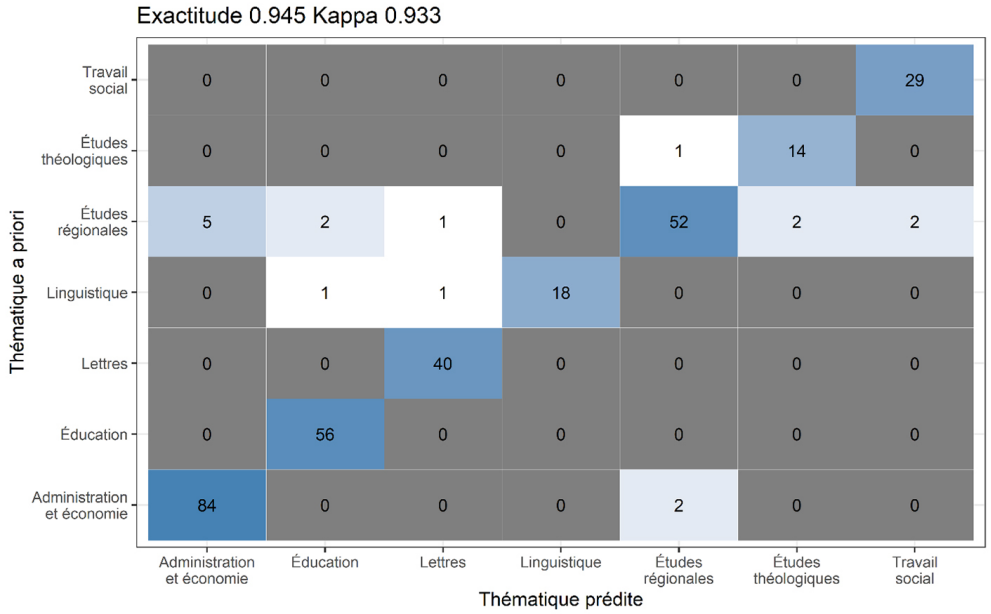


Figure 5. Matrice de confusion présentant le résultat de la tâche de classification automatique appliquée aux ressources textuelles de l'échantillon de test (n =310)[20]

19. Pour une description détaillée de ces hyperparamètres et de l'ensemble des hyperparamètres disponibles, le lectorat pourra consulter la documentation officielle à l'adresse suivante : https://xgboost.readthedocs.io/en/latest/

20. Le dégradé de couleurs utilisé dans cette figure illustre le degré d'adéquation entre la thématique *a priori* et la thématique prédite. Ce dégradé va de pair avec les informations chiffrées indiquées dans chaque cellule.

34. Les thèses et mémoires étiquetés en éducation, en travail social et en lettres ne font apparaître aucune confusion avec une autre thématique, leur taux de classification correcte est donc de 100 %. Les plus hauts degrés de confusion, bien que modestes, concernent les catégories études régionales, administration et économie. Ceci constitue une situation attendue puisque, par exemple, plusieurs maîtrises et doctorats en gestion de projet portent sur des questions régionales et inversement plusieurs études catégorisées comme régionales concernent des questions économiques. Il est donc logique que cet état de fait se reflète au sein même du vocabulaire utilisé. Notons aussi le faible chevauchement des catégories éducation, linguistique et lettres rendant compte d'une indépendance de ces trois thématiques de recherche.

35. Les taux de classification correcte des ressources issues de l'échantillon de test sont donc très élevés, ce qui signifie que l'algorithme présente de très bonnes performances. Différentes métriques peuvent être calculées pour estimer les performances d'un modèle de classification. À cet égard, l'exactitude (*accuracy*) est de 0,94. Il faut cependant rester prudent concernant l'utilisation de cette métrique dans le cadre d'une tâche de classification multiclasses qui, comme ici, présente des déséquilibres d'effectif entre les classes. Aussi, nous avons pris soin de calculer une seconde métrique, plus adaptée, le Kappa de Cohen (ou κ de Cohen) qui s'élève à $\kappa = 0{,}93$. Plus la valeur de ces deux métriques se rapproche de 1, plus un modèle de classification est considéré comme performant (pour une description de ces métriques, voir Grandini *et al.*, 2020).

36. L'algorithme *XGBoost* est de nature stochastique, les résultats pouvant donc varier d'une itération à l'autre de l'exécution informatique de l'algorithme. Cependant, la commande *set.seed*() utilisée à plusieurs reprises dans le *notebook* offre l'avantage pour l'utilisateur d'obtenir un même résultat à chaque itération du code source, peu importe la session de travail ou le système d'exploitation, garantissant ainsi une reproductibilité des résultats numériques proposés.

37. L'algorithme *XGBoost* offre aussi la possibilité de calculer l'importance relative des lemmes dans la prédiction de la classe d'appartenance d'un document. L'importance relative d'une variable explicative (*feature*) peut être mesurée par plusieurs métriques. Dans cette expérience, l'importance d'un lemme est fondée sur la notion de gain. Le gain représente la contribution relative d'une variable explicative au modèle. Une valeur plus élevée de cette métrique pour une variable explicative par rapport à une autre suggère que celle-ci est plus importante lors du calcul des prédictions.

38. Dans la figure 6 sont indiqués les quatre lemmes ayant la plus haute importance relative dans la prédiction de chacune de sept thématiques considérées. Là encore, étant donné la nature stochastique de l'algorithme *XGBoost*, nous avons choisi, comme indiqué dans le *notebook*, de calculer le gain moyen de chaque lemme après 100 itérations. Les

lemmes ayant le gain moyen le plus élevé dans chaque thématique reflètent indubitablement le contenu de ces dernières, tout en dessinant les contours des champs lexicaux abordés. Une fois encore, le code informatique utilisé pour construire cette analyse et la figure associée sont disponibles dans le *notebook*.

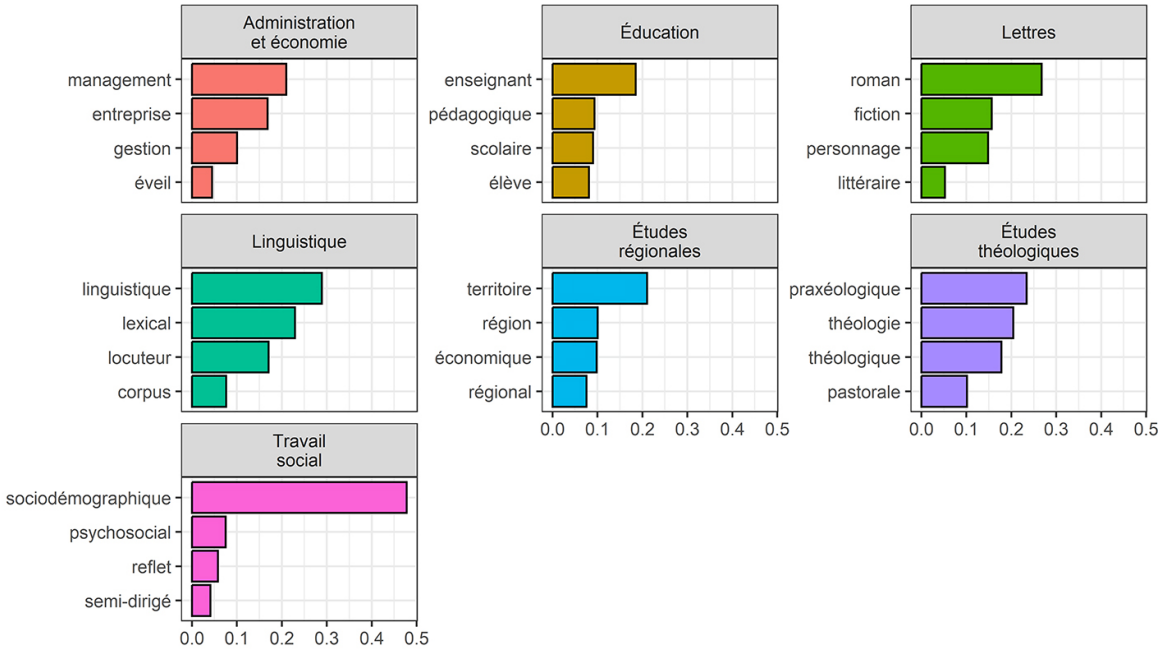


Figure 6. Importance relative des lemmes par thématique (après 100 itérations)

Conclusion

39. Le projet présenté dans ce chapitre s'inscrit dans le domaine pointu, mais éminemment d'actualité de l'apprentissage machine en traitement automatique des langues. Dans un tel domaine, les détails méthodologiques – nous avons par exemple évoqué ci-dessus l'optimisation des hyperparamètres – revêtent une importance fondamentale, car ils ont une influence certaine sur le résultat final de l'expérience. Sans pouvoir accéder aux détails techniques, la reproduction ou la réplication de la présente expérience serait quasiment impossible. Par conséquent, offrir un accès aux choix méthodologiques sous la forme d'un data paper, même si certaines notions peuvent échapper à un public non expert, est donc essentiel pour garantir les principes fondateurs de la science ouverte.

40. Ainsi, l'exemple de classification automatique de textes proposée dans ce chapitre ne peut être pleinement interprété qu'à la lumière des informations incluses dans le *notebook* disponible en ligne[21]. Les avantages de la pratique épistémologique proposée ici sont nombreux. Comme illustré dans le *notebook*, elle permet de rendre compte de l'exploration d'un échantillon et d'expliciter les analyses utilisées. Elle offre aussi la possibilité de préciser les choix méthodologiques et les résultats obtenus, tout en favorisant le partage des

21. https://doi.org/10.34847/nkl.ad3b77i7

données et la réutilisation du code informatique. À cet égard, on pourra noter que même les versions des bibliothèques logicielles utilisées ont été indiquées. Elle présente aussi des atouts pédagogiques certains en rendant explicite la séquentialité du processus d'analyse, tout en diminuant les coûts associés à la fouille, à l'analyse et au réemploi des données. En dernier lieu, elle soutient le processus d'évaluation des publications scientifiques, puisqu'elle rend disponibles les données utilisées et le détail de la méthodologie adoptée.

Remerciements

41. Les auteurs tiennent à remercier le Conseil de recherche en sciences humaines du Canada (CRSH) pour son soutien financier, la bibliothèque Paul-Émile Boulet de l'UQAC pour son soutien logistique, Johanna-Pascale Roy et Christophe Coupé pour leur relecture attentive du manuscrit et deux évaluateurs anonymes pour leurs suggestions.

42. La direction de l'ouvrage remercie Laurent Beauguitte et Robin Cura (UMR Geographie-cites, France) pour leur relecture.

Une analyse des modèles et instructions des *data papers* : types d'informations contextuelles décrites par les *data journals*

Jihyun Kim

Introduction

1. Le partage des données est une pratique émergente de la communication savante qui facilite le progrès de la science en rendant les données accessibles, vérifiables et reproductibles (Rowhani-Farid *et al.*, 2017). Il existe plusieurs façons de partager des données, notamment l'échange personnel de jeux de données, la publication de données sur les sites Web des chercheurs ou des laboratoires et le dépôt de données dans des entrepôts. Un moyen relativement nouveau de diffuser des ensembles de données est la publication de *data papers*, qui décrivent comment les données ont été collectées, traitées et vérifiées, améliorant ainsi la source des données (Pasquetto *et al.*, 2017). Les *data papers* sont publiés par des revues de données (*data journals*), et le processus de publication est similaire à celui des revues classiques, en ce sens que les *data papers* et les données sont tous deux évalués par des pairs, modifiés et rendus accessibles au public avec des identifiants uniques et pérennes (Gray, 2020). Comme les *data papers* prennent la forme de documents scientifiques et peuvent être cités par des articles de recherche, les créateurs des données peuvent être crédités à titre d'auteurs (Jefferies *et al.*, 2019).

2. Les *data papers* contiennent des faits sur les données plutôt que des hypothèses et des arguments résultant d'une analyse des données, comme c'est le cas dans les articles de recherche traditionnels (Chavan et Penev, 2011). Leur objectif principal est donc d'expliquer les données en fournissant « des informations sur le quoi, le où, le pourquoi, le comment et le qui des données » (Callaghan *et al.*, 2012). Le principal avantage des *data papers* est la richesse de leur documentation, qui est essentielle pour la réutilisation des données. Un *data paper* est généralement court et se compose d'un résumé, de méthodes de collecte et d'une description des données pertinents (Kratz et Strasser, 2014).

3. Cependant, des études antérieures ont mis en évidence l'absence de modèles ou d'instructions aux auteurs communs aux revues de données pour les *data papers*. Candela *et al.* (2015) ont identifié dix catégories de composants de *data papers* recommandés par les revues de données : accessibilité, conflit d'intérêts, domaine, format, licence, microattribution, projet, provenance, qualité et réutilisation. Les auteurs ont noté qu'un identifiant unique indiquant la disponibilité des données, tel qu'un DOI ou un URI, était la seule information fournie

par tous les *data journals* qu'ils ont étudiées. Moins de la moitié des revues de données demandaient des informations sur le domaine, la licence, la micro-attribution, le projet et la réutilisation (Candela *et al.*, 2015). En outre, Chen a expliqué que les modèles/instructions des *data papers* se concentrent principalement sur des jeux de données uniques, c'est-à-dire au niveau de l'élément, et que seuls quelques-uns fournissent des descriptions de données au niveau de la collection, comme des jeux de données multiples ou des bases de données. L'auteur a suggéré que la granularité des données de recherche qu'un *data paper* décrit soit spécifiée par les revues de données (Chen, 2017).

4. Le manque de normalisation et le problème de la granularité dans la description des données ont été abordés dans d'autres études concernant la documentation des données et les métadonnées (Atici *et al.*, 2013 ; Friedhoff *et al.*, 2013 ; Kim *et al.*, 2019). Ces études ont également indiqué que la documentation d'une quantité adéquate d'informations contextuelles appropriées sur les données augmenterait le potentiel de réutilisation des données. De même, l'objectif sous-jacent de la publication de *data papers* est de permettre la réutilisation des données, et l'on s'attend à ce que ces articles traitent des défis qui entraînent un « échec de la réutilisation des données » (Rees, 2010 ; Costello *et al.*, 2013). Dans ce contexte, la présente étude examine les types d'informations contextuelles que les revues de données demandent de décrire et détermine dans quelle mesure ces types d'informations sont communs ou variables selon les revues.

5. Cette étude vise à identifier les composants d'un *data paper* tels que définis par les modèles et instructions aux auteurs de 24 revues de données indexées par *Web of Science* (WoS), le type de document étant limité aux *data papers*. Les composants des *data papers* ont été mis en correspondance avec les types d'informations contextuelles suggérés par des études antérieures (Faniel *et al.*, 2019 ; Chin et Lansing, 2004). Par conséquent, il est possible de déterminer dans quelle mesure les *data papers* publiés dans diverses revues couvrent les informations contextuelles dont les chercheurs ont besoin pour la réutilisation des données et d'identifier les éléments communs et uniques dans les revues de données. Les résultats visent à aider les chercheurs à mieux comprendre comment améliorer les instructions aux auteurs des revues de données pour documenter les données et le rôle même des revues de données.

Méthodologie

6. Cette étude a d'abord identifié un large ensemble de *data papers* sur la base 1) de deux études (Candela *et al.*, 2015 ; Chen, 2017) qui ont réalisé une analyse de contenu de modèles et/ou d'instructions de *data papers* et 2) d'une liste de *data papers* rapportée par Akers (2004). Candela *et al.* (2015) ont analysé 116 revues de données de 15 éditeurs par le biais de recherches sur Internet. Chen (2017) a créé une liste initiale de 93 revues de données sur la base de la liste d'Akers et de recherches sur UlrichsWeb et a sélectionné 26 revues de données de 16 éditeurs en tenant

compte des domaines disciplinaires. En excluant les revues en double des deux études, on a obtenu 106 revues de données. Comme les études précédentes (Candela *et al.*, 2015 ; Chen, 2017) l'ont suggéré, la grande majorité des revues de données étaient des revues mixtes (c'est-à-dire des revues publiant tout type d'article, y compris des *data papers*), et les revues de données pures (c'est-à-dire les revues publiant uniquement des *data papers*) ne représentaient qu'une petite proportion.

7. Cette étude a utilisé le WoS comme outil pour sélectionner les revues de données pour l'analyse. Malgré les débats sur la fiabilité des facteurs d'impact des revues générés par le WoS, les revues indexées par le WoS conservent généralement un bon statut, car elles doivent répondre aux critères de qualité fixés par la base de données. Sur les 106 revues consacrées aux données, 79 étaient indexées par le WoS. La recherche a été limitée au type de document « *data papers* » dans la fonction de recherche avancée de la base de données, et 24 revues de données ont finalement été sélectionnées (Annexe 1) (Candela *et al.*, 2015 ; Chen, 2017). Dix-huit de ces 24 revues se recoupent avec celles examinées par l'une ou l'autre ou les deux études susmentionnées (10 par Candela *et al.* (2015), 1 par Chen (2017) et 7 par les deux). Les six autres revues, à savoir *BioInvasions Records*, *Data*, *Ecological Research*, *Journal of Hymenoptera Research*, *Frontiers in Marine Science*, et *Comparative Cytogenetics*, ont également été analysées dans cette étude.

8. Sur les 24 revues de données, sept ont été publiées par Springer Nature et six par Pensoft (annexe 1). Toutes les revues publiées par Pensoft ont utilisé les mêmes instructions aux auteurs pour les *data papers*. Parmi les sept revues de Springer Nature, cinq revues de BioMed Central (BMC) ont partagé les mêmes instructions aux auteurs pour les *data papers*. Les autres revues de données ont fourni leurs propres modèles et/ou instructions pour les *data papers*. Par conséquent, cette étude a rassemblé 15 modèles et/ou instructions de *data papers* distincts pour l'analyse.

9. Pour étudier les informations contextuelles couvertes par les *data papers*, j'ai utilisé les types d'informations contextuelles suggérés par Faniel *et al.* (2019) et Chin et Lansing (2004), qui ont élaboré une série de contextes de données reflétant les perspectives des réutilisateurs de données. Chin et Lansing (2004) ont proposé à l'origine divers attributs de contextes scientifiques et sociaux qui facilitent le partage des données dans les collaborations scientifiques biologiques. Quatre de ces attributs contextuels sont particulièrement pertinents pour le contexte scientifique, et ont donc été utilisés pour l'analyse (tableau 1) (Candela *et al.*, 2015 ; Chen, 2017 ; Faniel *et al.*, 2019 ; Chin et Lansing, 2004). J'ai ensuite mis en correspondance les types d'informations contextuelles avec les composantes des *data papers* identifiés par Candela *et al.* (2015) et Chen (2017). Cette mise en correspondance a permis une évaluation préliminaire de la relation entre les composantes des *data papers* et les informations contextuelles et le développement d'un schéma de codage pour l'analyse de contenu.

	Types of contextual information		Data paper components	
	Faniel et al. [15]	Chin and Lansing [16]	Candela et al. [8]	Chen [9]
Data production information	Data collection	Experimental properties	Quality, provenance	Collection
	Specimen and artifact	–	Coverage	Coverage
	Data producer	–	Microattribution	Description (file creators), author's contribution
	Data analysis	Analysis and interpretation	–	–
	Missing data	–	–	–
	Research objectives	–	–	–
Repository information	Provenance	Data provenance	Availability	Identifier, relationship
	Repository reputation and history	–	–	–
	Curation and digitization	–	–	–
Data reuse information	Prior reuse	–	Reuse	
	Advice on reuse	–	Reuse	–
	Terms of use	–	License, competing interests	Copyright, ethical approval, consent to publication, competing interests
–	–	General data set properties	Format	Description (file format, version, creation date)
–	–	–	Project	Funding statement

Tableau 1. Correspondance entre les types d'informations contextuelles et les composantes des *data papers*

10. Le tableau 1 montre que les types d'informations contextuelles suggérés par Faniel *et al.* (2019) ne correspondent pas tous aux composants des *data papers*. Plus précisément, aucune composante des *data papers* identifiée par les études précédentes ne correspondait à « l'analyse des données », « les données manquantes », « les objectifs de la recherche », « la réputation et l'historique du dépôt » et « la conservation et la numérisation ». La définition du terme « provenance » n'est pas non plus uniforme d'une étude à l'autre. Candela *et al.* (2015) ont défini cette notion comme la « méthodologie menant à la production de l'ensemble de données », ce qui est plus proche de la « collecte de données » que la définition de Faniel *et al.* comme « sources du matériel ou traçabilité ».

11. « *General data set properties* », proposé par Chin et Lansing (2004), correspondait à une composante des *data papers* relative à la description des formats, versions et dates de création des données. En outre, le « projet », mentionné par Candela *et al.* (2015), qui fait référence aux

informations sur les initiatives dans le cadre desquelles les données sont générées, était le seul composant qui ne correspondait à aucun des types d'informations contextuelles. La « déclaration de financement », identifiée par Chen (2017), était également liée à l'élément « projet ». Connaître les informations sur un projet et les sources de financement qui ont conduit à la création des données serait utile lorsqu'on envisage la possibilité de réutiliser les données. Ainsi, cette étude a considéré les informations sur le projet comme des informations contextuelles supplémentaires.

12. Le schéma de codage pour l'analyse des composantes des *data papers* était largement basé sur les types d'informations contextuelles suggérés par Faniel *et al.* (2019). En outre, la composante « propriétés générales du jeu de données », qui a été notée par Chin et Lansing (2004), a été ajoutée au schéma de codage. La composante « projet » a été incluse dans la catégorie « informations sur la production des données », proposée par Faniel *et al.* (2019), car il s'agit d'un facteur contextuel pertinent pour la création des données.

Résultats

13. Les types d'informations contextuelles examinés dans cette étude sont classés en quatre groupes : propriétés générales des jeux de données, informations sur la production des données, informations sur l'entrepôt et informations sur la réutilisation des données. En ce qui concerne les propriétés générales des ensembles de données, l'étude cherche à savoir si les 15 modèles et/ou instructions aux auteurs pour les *data papers* exigent des auteurs qu'ils décrivent les attributs énumérés dans le tableau 2. Les dates de création, les formats et les versions des données sont mentionnés par Chen (2017), et les autres propriétés sont identifiées au cours du processus de codage.

General data set properties	J_{BMC} [a]	SD	BMC genet	$J_{neuroft}$ [b]	ER	Ecol	GDJ	ESSD	DiB	FiMS	Data	GigaSci	BIR	IJRR	JOAD
Data creation date															√
Data format		√	√		√	√	√		√						√
Data type							√		√			√			√
Data version				√			√								√
File name/title			√	√		√	√			√	√	√			√
File size						√		√				√			
Language				√											√

Tableau 2. Propriétés générales des jeux de données identifiées dans les modèles/instructions pour les *data papers*

14. Le nom/titre du fichier de données est la seule propriété demandée par huit des modèles/instructions (53,3 %) ; la description du format des données est quant à elle

demandée par sept modèles (46,7 %). Les autres propriétés de l'ensemble de données ne sont pas souvent requises par les revues de données : les descriptions des dates de création des données et des langues sont rares. Le type de données est défini différemment selon les revues ; par exemple, *Journal of Open Archaeology Data* (*JOAD*) fait la distinction entre les données primaires, les données secondaires, les données traitées, l'interprétation des données et les rapports finaux, tandis que *Data in Brief* (*DiB*) classe les données par type : tableaux, images, diagrammes, graphiques et figures.

15. Les informations sur la production des données ont tendance à être demandées plus fréquemment par les revues de données que les propriétés générales des jeux de données (tableau 3). Tous les modèles et toutes les instructions aux auteurs exigent des informations relatives à la collecte des données, principalement en ce qui concerne les étapes de la collecte des données, les stratégies d'échantillonnage et les mécanismes de contrôle de la qualité. Des descriptions des producteurs de données sont requises par neuf modèles/instructions, dont cinq demandent spécifiquement des informations sur les créateurs de données (liste des auteurs des ensembles de données [*GigaScience*] ou des créateurs [*Jpensoft*, *Geoscience Data Journal*, *Data*, *JOAD*]). Les quatre autres revues demandent une description des contributions ou des informations des auteurs, ce qui correspond peut-être au rôle de créateur de données. La composante « projet » est mentionnée par sept modèles/instructions, et seuls deux d'entre eux exigent une description globale du projet (*Ecology* et *Jpensoft*). Les cinq autres revues exigent des informations sur le financement.

Data production information	J_{BMC} [a]	SD	BMC genet	$J_{pensoft}$ [b]	ER	Ecol	GDJ	ESSD	DiB	FiMS	Data	GigaSci	BIR	IJRR	JOAD
Data collection	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Specimens and artifacts				√				√	√			√			√
Data producer	√	√	√	√			√	√			√	√			√
Data analysis		√	√			√				√					
Missing data			√		√	√					√			√	
Research objectives		√	√			√	√					√			
Project	√		√	√		√		√				√			√

Tableau 3. Informations relatives à la production de données identifiées dans les modèles/instructions des *data papers*

16. Les informations sur les spécimens et les artefacts sont demandées par cinq modèles/instructions de revues de sciences biologiques, de géosciences et d'archéologie, dont les chercheurs de ces disciplines ont besoin pour la réutilisation des données (Faniel *et al.*, 2019). La couverture temporelle, spatiale ou taxonomique (*Jpensoft*, *JOAD*), la disponibilité ou l'emplacement des échantillons (*Earth System Science Data*, *DiB*) et les descriptions des organismes ou des tissus (*GigaScience*) sont identifiés. Des informations sur l'analyse des données, les données manquantes

et les objectifs de recherche sont également demandées par quatre ou cinq modèles/instructions. L'information sur l'analyse des données porte sur la façon dont les données sont traitées, et l'information sur les données manquantes porte sur les anomalies ou le bruit des données. Les objectifs de recherche sont souvent exprimés en tant que motivations ou justifications de la collecte des ensembles de données.

Types of contextual information	J_{BMC} [a)]	SD	BMC genet	$J_{neurosoft}$ [b)]	ER	Ecol	GDJ	ESSD	DiB	FiMS	Data	GigaSci	BIR	IJRR	JOAD
Repository information															
Provenance	√	√	√	√	√	√	√	√	√	√	√	√		√	√
Repository reputation and history															
Curation and digitization						√									
Data reuse information															
Prior reuse	√					√	√								
Advice on reuse	√	√					√		√	√	√	√			√
Terms of use	√	√	√		√	√	√	√	√		√	√			√

Tableau 4. Informations relatives au dépôt et à la réutilisation des données identifiées dans les modèles/instructions pour les *data papers*

17. En termes d'informations sur le dépôt, toutes les revues sauf une demandent de décrire la provenance des données, en indiquant l'identifiant ou l'emplacement des données (tableau 4). La provenance des données fait également référence à la relation des données avec d'autres matériaux, bien qu'une seule revue (*DiB*) demande une description de tout article de recherche lié aux données. Aucune des revues sur les données ne demande de description sur la réputation et l'historique de l'entrepôt. En ce qui concerne la conservation et la numérisation, une revue (*Ecology*) demande des informations sur les procédures d'archivage, y compris une description de la manière dont les données sont archivées pour un stockage et un accès pérenne.

18. Des informations sur la réutilisation des données, concernant principalement des conseils sur la réutilisation et les conditions d'utilisation, sont demandées (tableau 4). En ce qui concerne les conseils sur la réutilisation, les auteurs doivent décrire la réutilisation potentielle et l'intérêt de leurs données pour la réutilisation. Les revues de données demandent aux auteurs de décrire les conditions d'utilisation, principalement en ce qui concerne les intérêts concurrents, mais aussi plusieurs autres aspects, notamment l'approbation éthique et le consentement à la participation, le consentement à la publication, la licence, le droit d'auteur et les exigences d'accessibilité.

Discussion

19. Les résultats ont révélé des types communs et uniques d'informations contextuelles que les revues de données

demandent aux auteurs de décrire. La forme la plus courante d'informations contextuelles documentées par les revues concerne les méthodes de collecte des données, suivies de la provenance des données (emplacement des dépôts et/ou identifiants des données). Plus de la moitié, ou presque, des modèles/instructions identifient les noms/titres des fichiers de données et les formats de données comme propriétés générales des jeux de données, les informations sur la production des données (y compris le producteur de données et le projet) et les informations sur la réutilisation (y compris les conseils sur la réutilisation et les conditions d'utilisation). Les résultats sont pour la plupart conformes à ceux d'études antérieures (Candela *et al.*, 2015 ; Chen, 2017). Pourtant, si Candela *et al.* (2015) mentionnent que les descriptions des informations de réutilisation sont souvent négligées par les revues de données, la plupart des revues de données examinées dans cette étude abordent la réutilisation potentielle des données. En termes de provenance des données (indiquant la relation entre les données et d'autres objets), seule une revue (*DiB*) dans cette étude exige des informations sur cette relation, bien que Chen (2017) ait identifié plus de cas où ces informations sont requises.

20. Les types d'informations contextuelles que les revues de données ne demandent jamais ou rarement comprennent des informations sur l'entrepôt (réputation et historique de l'entrepôt, conservation et numérisation) et les propriétés des jeux de données (dates de création des données et langues). En particulier, Faniel *et al.* (2019) ont déclaré que la réputation et l'historique de l'entrepôt sont moins faciles à documenter, car ils sont plus sociaux et relatifs que d'autres types de contexte. Deux des revues de données examinées dans cette étude (*Scientific Data* et *Earth System Science Data*) fournissent des critères pour recommander des entrepôts de données, à savoir les conditions d'accès aux données et la disponibilité à long terme[1][2]. Ces informations fournies par les revues de données sur les entrepôts visent à aider les réutilisateurs à comprendre la fiabilité des entrepôts où certaines données sont déposées. Alors que les dates de création des données aideront les réutilisateurs à développer des cadres d'échantillonnage et à identifier les changements dans les contextes de création des données (Zimmerman, 2007), seule une revue demande cette information.

21. Les informations sur la production des données concernant l'analyse des données, les données manquantes et les objectifs de la recherche n'ont pas été précisées par les études de Candela *et al.* (2015) et de Chen (2017) (tableau 1). Cependant, quatre ou cinq des modèles/instructions demandent une description de ces informations. En outre, les revues de données demandent rarement des informations sur la version des données, la taille du fichier et la réutilisation antérieure, que trois des modèles/instructions mentionnent.

1. *Scientific Data*. Suggesting additional repositories. *Nature Research* https://www.nature.com/sdata/policies/data-policies#repo-suggest
2. *Earth System Science Data*. Repository criteria. https://www.earth-system-science-data.net/for_authors/repository_criteria.html

22. Dans l'ensemble, seule une petite quantité d'informations contextuelles est couramment demandée par les revues de données. Les instructions ont tendance à se concentrer davantage sur les informations relatives à la production des données (collecte des données, producteur de données et projet) et sur les informations relatives à la réutilisation (réutilisation potentielle et conditions d'utilisation) que sur les propriétés générales des jeux de données ou les informations relatives à l'entrepôt. À l'exception des noms de fichiers et des formats de données, les descriptions des propriétés des ensembles de données sont généralement absentes. Les informations sur les entrepôts de données concernent principalement les identifiants uniques des données, mais les revues de données pourraient fournir des informations sur la réputation des entrepôts ou leurs pratiques de conservation afin d'aider les lecteurs des *data papers* à évaluer la réutilisabilité des données.

Conclusion

23. En conclusion, la présente étude examine les types d'informations contextuelles que les revues de données demandent aux auteurs de décrire et détermine l'ampleur de la variation de ces informations selon les revues de données. La principale motivation de la publication de *data papers* est de rendre les données réutilisables et reproductibles et les *data papers* devraient fournir une documentation étendue sur les données qui reflète suffisamment d'informations contextuelles. Cette étude suggère que les revues de données auraient intérêt à fournir un ensemble plus standardisé des composantes des *data papers* afin d'informer les réutilisateurs des différents types d'informations contextuelles de manière cohérente. En outre, les revues de données devraient non seulement exiger des informations sur la disponibilité des données, mais aussi de fournir des détails sur la qualité des entrepôts de données qui viendraient compléter les informations sur les entrepôts décrites par les auteurs de *data papers*.

Informations sur l'article

Citation de l'œuvre originale

26. Kim, J. (2020). An analysis of data paper templates and guidelines: types of contextual information described by data journals. *Science Editing*, *7*(1), 16–23. https://doi.org/10.6087/kcse.185, https://www.escienceediting.org/journal/view.php?viewtype=pubreader&number=201, https://doi.org/10.6087/kcse.185

Annexes

Annexe 1. Vingt-quatre *data journals* sélectionnés pour l'analyse

Publisher	Data journal (full name)	Subject area	Publishing model	Pure vs. mixed
Springer Nature	BMC Bioinformatics[a)]	Biochemical research methods	OA	Mixed
	BMC Genetics	Genetics, heredity	OA	Mixed
	BMC Genomics[a)]	Biotechnology, applied microbiology	OA	Mixed
	BMC Medical Genomics[a)]	Genetics, heredity	OA	Mixed
	BMC Medical Informatics and Decision Making[a)]	Medical informatics	OA	Mixed
	BMC Musculoskeletal Disorders[a)]	Orthopedics	OA	Mixed
	Scientific Data	Multidisciplinary sciences	OA	Pure
Pensoft[b)]	Biodiversity Data Journal	Biodiversity conservation	OA	Mixed
	Comparative Cytogenetics	Genetics, heredity	OA	Mixed
	Journal of Hymenoptera Research	Entomology	OA	Mixed
	Neobiota	Biodiversity conservation	OA	Mixed
	Phytokeys	Plant sciences	OA	Mixed
	Zookeys	Zoology	OA	Mixed
Wiley	Ecological Research	Ecology	Hybrid	Mixed
	Ecology	Ecology	Hybrid	Mixed
	Geoscience Data Journal	Geosciences, multidisciplinary	OA	Pure
Copernicus Publications	Earth System Science Data	Geosciences, multidisciplinary	OA	Pure
Elsevier	Data in Brief	Multidisciplinary sciences	OA	Pure
Frontiers Media S.A.	Frontiers in Marine Science	Environmental sciences	OA	Mixed
MDPI	Data	Computer science information systems	OA	Mixed
Oxford University Press	GigaScience	Multidisciplinary sciences	OA	Pure
REABIC Journals	BioInvasions Records	Biodiversity conservation	OA	Mixed
SAGE Publications	International Journal of Robotics Research	Robotics	Subscribe	Mixed
Ubiquity Press	Journal of Open Archaeology Data	Archaeology	OA	Pure

Annexe 2. *Data journals* utilisant le même *guideline* inclus dans l'analyse

Éditeur	Data journals
JBMC	BMC Bioinformatics
	BMC Genomics
	BMC Medical Genomics
	BMC Medical Informatics and Decision Making
	BMC Musculoskeletal Disorders
Jpsensoft	Biodiversity Data Journal
	Comparative Cytogenetics
	Journal of Hymenoptera Research
	Neobiota
	Phytokeys
	Zookeys

Qualité, évaluation, compétences

Évaluer un *data paper*, l'exemple de *Cybergeo*

Clémentine Cottineau-Mugadza, Christine Kosmopoulos et Denise Pumain

Introduction

1. Les *data papers* ont une fonction essentielle dans la valorisation et la préservation du savoir scientifique. Leur écriture se situe au plus près du travail de recherche, car, la plupart du temps, ces données résultent de l'important travail engagé par les auteurs pour obtenir des résultats publiables, soit que les données aient été directement collectées par les chercheurs, soit qu'elles aient subi un traitement spécifique, en principe conçu de façon optimale pour résoudre une question de recherche bien précise. Le *data paper* permet non seulement de partager ces données avec d'autres chercheurs ou utilisateurs, mais aussi de préserver la mémoire de leur construction, telle qu'elle a été imaginée et réalisée selon la problématique des auteurs. Cette restitution des informations a donc une double utilité, autant pour les auteurs de l'article que pour ses lecteurs. Le *data paper* décrit des données scientifiques et les circonstances et méthodes de leur collecte, il aide à comprendre leur contenu et leurs limitations pour en faciliter la réutilisation, il donne une visibilité aux données et assure leur citabilité, même avant que l'analyse des données soit terminée et/ou publiée. Une partie non négligeable des critères à prendre en compte pour l'évaluation de ces articles est déterminée par ce premier ensemble d'objectifs. Ce sujet est encore peu traité dans la littérature scientifique : une recherche rapide sur *Google Scholar* faite en avril 2022 n'identifie que quelques rares publications sur la qualité des données associées aux *data papers* (par exemple sur le sujet de la biodiversité Egloff *et al.* 2016) et ne trouve que deux articles en réponse au mot-clé « *evaluation of data papers* », l'un en géomorphologie expérimentale (Hsu *et al.*, 2015) et l'autre en épistémologie des sciences (Li et Jiao, 2021).

2. Le processus d'évaluation souscrit également aux critères habituels de l'évaluation scientifique dans le domaine concerné. La publication de *data papers* s'inscrit plus généralement dans la démarche d'une science reproductible, qui produit des savoirs partageables et cumulables. Une telle exigence est bien entendu compatible avec la conscience, toujours présente parmi les scientifiques, que les contenus de ces savoirs sont périodiquement révisables à la lumière d'interprétations reformulées, ou d'éclairages nouveaux. Mais ces transformations des connaissances scientifiques ne doivent pas cautionner une posture relativiste selon laquelle les savoirs seraient comparables à des opinions, et variables selon les personnes et selon les moments. À chaque moment, il est possible de dresser un état des connaissances qui s'appuie sur un large consensus parmi les spécialistes

d'une question scientifique donnée. Les révisions de cet état peuvent être apportées par de nouvelles données, par l'ouverture des corpus d'analyse à des cas non encore intégrés, et par l'introduction de nouveaux instruments d'observation, de mesure ou de mise en relation, ou par de nouvelles méthodes de validation. Les consensus sur un état de connaissances s'établissent entre les scientifiques par discussion entre les pairs, qui s'accordent sur des énoncés de propositions théoriques, des procédures et méthodes d'analyse adaptées et des procédures de probation reconnues. L'évaluation d'un *data paper* doit aussi tenir compte de ces critères habituels appliqués à l'évaluation d'un article scientifique.

3. En sciences humaines et sociales, l'évaluation des publications scientifiques par les pairs peut être encore plus compliquée du fait de la pluralité des théories et des approches courantes, et de la diversité et complexité des objets de recherche, sans doute plus grandes que celles des objets étudiés par les sciences de la matière ou même de celles des sciences de la vie (Kosmopoulos et Dassa, 2011). En conséquence, il n'existe pas de modèle standard de construction et d'écriture des articles, comme on peut les trouver en biologie ou en physique. Les évaluations s'appuient donc sur un large ensemble de paramètres, qui précisent les références théoriques des articles aussi bien que le contexte spatio-temporel des cas étudiés pour ces « sciences historiques » (Passeron, 2011). Les *data papers* doivent être évalués aussi selon ces critères, même s'ils sont construits selon des normes relativement contraignantes pour permettre la compréhension et la réutilisation des données qu'ils apportent. Nous rappelons brièvement en premier lieu (section 1) quels sont ces critères généraux d'évaluation. Nous détaillons dans une seconde partie l'ensemble de critères qui doivent être partagés par les auteurs et les évaluateurs pour assurer la cohérence et la validité d'un *data paper* dans le cas particulier d'une revue de géographie qui reçoit des données très spécifiques (section 2). Nous montrons enfin quelles adaptations ont été nécessaires au niveau de l'organisation d'une revue pour publier ce type d'articles, en nous appuyant sur le cas précis de la revue en ligne *Cybergeo, revue européenne de géographie* (section 3).

Principes généraux de l'évaluation appliqués aux *data papers*

4. Les principes généraux de l'évaluation d'articles que nous appliquons dans *Cybergeo* sont de deux ordres. D'une part, l'article doit contribuer à la connaissance scientifique en proposant un contenu inédit, original et jugé digne d'intérêt pour la communauté disciplinaire à laquelle s'adresse la revue, comme c'est le cas pour les articles classiques. D'autre part, l'article doit démontrer rigoureusement la valeur des résultats apportant une contribution à la connaissance scientifique, en précisant notamment les données et méthodes utilisées pour produire lesdits résultats, leur adéquation avec la question de recherche, leur correspondance avec les autres recherches menées par d'autres scientifiques dans le domaine d'étude, etc.

5. Dans le cas d'un *data paper*, les résultats obtenus et proposés comme contribution à la connaissance scientifique sont des données. Il faut donc pour leurs auteurs justifier de l'intérêt de ces données pour la communauté disciplinaire à laquelle s'adresse la revue. En effet, des données sans vocation à être réutilisées présentent un intérêt mineur en termes de publication. Leur description et leur archivage tiennent de la bonne pratique scientifique, mais ne nécessitent pas le labeur additionnel de rédaction d'un *data paper*.

6. La démonstration de la valeur de ces données-résultats passe par la description des sources utilisées et des méthodes de collecte, en lien avec l'état de l'art scientifique. Les opérations sur les données – telles que le « nettoyage », l'harmonisation, l'estimation des données manquantes, leur agrégation/désagrégation, la création/transformation de nouvelles variables, etc. – doivent être détaillées de façon à ce qu'un lecteur attentif soit en mesure de reproduire ces opérations avec le même résultat, être justifiées par la question de recherche, et leur conséquence évaluée en termes absolus (quelles sont les conséquences sur les données finales ?) et relatifs (en quoi sont-elles comparables aux pratiques des recherches menées par d'autres scientifiques dans le domaine d'étude ?).

7. Il est d'usage de faire évaluer les articles scientifiques par plusieurs spécialistes de l'objet étudié dans les articles. Dans le cas d'un *data paper*, l'objet d'étude est double et requiert deux types de compétences de la part des évaluateurs : des connaissances sur la production de données (connaissance des sources, des méthodes et techniques de production, de gestion et de transformation de données) afin d'évaluer la rigueur et l'opportunité des opérations décrites, et des connaissances sur le thème des données, afin d'évaluer l'intérêt du résultat et la justification des choix opérés. Cette double compétence peut être fournie par un même évaluateur, ou plus souvent par deux (ou plus) spécialistes distincts.

Critères d'évaluation dans le cas particulier d'une revue de géographie

8. Dans le cas de *Cybergeo*, 10 *data papers* ont été publiés entre la création de la rubrique en 2017 et 2021 (13 en 2024). Ils sont soumis à la procédure habituelle d'évaluation de la revue, qui se pratique en double aveugle[1]. Les auteurs soumettent leur article sur la plateforme en choisissant la rubrique Data Papers. Un comité de pré-sélection examine rapidement l'article et décide s'il présente assez d'intérêt et de qualité de présentation formelle pour être soumis aux lecteurs de la revue. À ce stade on vérifie aussi que l'article est convenablement anonymisé, on s'assure de sa non-publication antérieure et de l'absence de plagiat. L'article est alors proposé au comité de lecture (voir *infra* section « Un nouveau comité de lecture » pour plus de précision). Au moins deux et souvent trois évaluateurs déposent leur avis sur la plateforme. Un membre du comité de direction de la

1. https://journals.openedition.org/cybergeo/23412

revue fait une synthèse de ces avis au nom du comité de rédaction, et cette synthèse est adressée aux auteurs, ainsi qu'aux personnes ayant donné leur avis pour l'évaluation. À aucun moment les évaluateurs n'ont connaissance des noms des auteurs, et, réciproquement, les auteurs ignorent par qui leur article a été évalué.

9. Les avis des évaluateurs sont inclus dans la synthèse, ils demandent le plus souvent des précisions aux auteurs et leur suggèrent des améliorations pour rendre l'article publiable. Une deuxième version ainsi révisée est alors déposée par les auteurs sur la plateforme. Selon l'ampleur des modifications demandées, cette nouvelle version est à nouveau soumise à évaluation et le processus peut être itéré, parfois jusqu'à une 4e ou 5e version, avant que l'article ne soit mis en ligne. La durée du processus d'évaluation d'un article posant peu de problèmes est de l'ordre de deux mois.

10. Les évaluateurs s'assurent de la qualité des articles en suivant un questionnaire précis contenant les critères suivants : originalité du sujet, clarté de l'exposition de la recherche et explication de la démarche adoptée, situation de la problématique dans le débat scientifique national et international, cohérence et suivi de l'argumentation, précision de la méthodologie, pertinence des références bibliographiques, qualité des illustrations… Les critères s'adaptent selon l'orientation choisie par les auteurs et selon l'état des connaissances dans le domaine. Pour les *data papers*, l'évaluation tient compte aussi de l'envergure des bases de données apportées et des enjeux de leurs possibles réutilisations, en lien avec les principes de la science ouverte préconisés en Europe à partir de 2016 sous l'appellation FAIR (Findable, Accessible, Interoperable, Reusable). Les bases de données ainsi évaluées par les pairs ont une grande valeur ajoutée par rapport aux simples entrepôts de données. Les ensembles de données deviennent compréhensibles, accessibles et explorables avec des moteurs de recherche sémantiques et géographiques.

11. Les *data papers* de *Cybergeo* sont plus spécifiques que leurs équivalents célèbres (*Nature's Scientific Data*, *Open Health Data*, etc.), car ils retiennent exclusivement des bases de données géographiques, c'est-à-dire des données incluant des informations géolocalisées caractérisant une portion de l'espace terrestre ou des sociétés qui le pratiquent. C'est pourquoi les recommandations aux auteurs demandent de préciser les échelles concernées, les composantes spatiales, la géométrie et la compatibilité avec les systèmes d'information géographique (SIG) des éléments inclus dans la base de données, en plus des informations requises habituellement telles que les sources originales, la méthode de construction des données et les procédures de validation.

Visibilité internationale et réutilisation maximale

12. L'équipe fondatrice de *Cybergeo* souscrit à l'idée que les sciences sociales doivent pouvoir être publiées dans la langue des chercheurs qui les produisent, pour beaucoup

de raisons qui tiennent aux spécificités des langues et des cultures et aux idiosyncrasies des agendas de recherche qui en résultent. C'est ainsi que, depuis sa création en 1996, *Cybergeo* accepte des articles rédigés dans chacune des principales langues européennes. Dans le cas des *data papers*, nous avons modifié cette exigence pour privilégier un accès élargi aux données et faciliter leur réutilisation. Le partage de ces informations implique d'employer un langage plus généralement répandu dans la pratique des scientifiques internationaux, et plus standardisable, qui facilite leur réutilisation : « Les *data papers* devant être le plus souvent possible rédigés en anglais pour une plus large diffusion internationale, les recommandations aux auteurs concernant cette rubrique sont rédigées dans cette langue »[2].

Reproductibilité

13. Les *data papers* publiés dans *Cybergeo* sont des articles évalués par les pairs qui sont destinés à décrire, documenter et évaluer des bases de données géographiques produites par des auteurs. L'article contient un lien vers les données qui sont stockées dans un entrepôt pérenne (de préférence l'entrepôt Dataverse de *Cybergeo*[3]) et mises à disposition sous licence libre (CC0-1.0, CC-BY-4.0 ou CC-BY-SA-4.0 selon la classification Creative Commons[4]). Il doit documenter, expliquer et justifier toutes les procédures utilisées pour transformer les enregistrements originaux produisant la géodatabase finale, démontrer sa nécessité et son utilisation, et guider le lecteur sur la manière de la réutiliser. Le lecteur, muni de ces informations et des certificats d'autorisation appropriés, devrait être en mesure de reproduire entièrement les résultats énoncés dans l'article. Le lecteur peut aussi réutiliser les données pour d'autres objectifs, en respectant les normes de citation demandées.

14. Des consignes spécifiques exigeant des précisions quant au format des métadonnées, l'origine des données, la datation, les institutions responsables, sont précisées dans une liste destinée aux auteurs. Il est en effet important pour la reproductibilité des résultats que l'article fournisse des détails sur toutes les sources des enregistrements originaux qui ont été utilisées pour construire la base de données (organisation, datation temporelle, format, variables sélectionnées, lien avec le référentiel, etc.).

Explicitation et précision de l'information géographique

15. Du fait de la particularité géographique des *data papers* publiés par *Cybergeo*, les caractéristiques spatiales et scalaires des données doivent être explicitées (points, lignes, polygones, local, régional, etc.). La lecture de l'article doit permettre de fournir une réponse argumentée aux questions suivantes : de quelles unités

2. https://journals.openedition.org/cybergeo/23412#tocto1n7
3. https://dataverse.harvard.edu/dataverse/cybergeo
4. https://creativecommons.org/publicdomain/zero/1.0/

spatiales les données finales sont-elles composées ? Sont-elles compatibles avec toute géométrie SIG disponible ? Comment les objets de sortie sont-ils liés aux objets d'entrée (agrégation, transformation, etc.) ? Les normes ISO sont employées, dans la mesure du possible, pour apporter toutes ces précisions (tableau 1).

Type of spatial representation			controlled list	vector - grid - table - tin - esteoroscopic model - photo - video
Spatial resolution (scale or minimum cartographic unit)			text	50 characters
Language			text	50 characters
Themes			text	100 characters
Geographic extension			text	
	min y		number - float	
	min x		number - float	
	max y		number - float	
	max x		number - float	
Reference system			text	50 words
Source			text	100 words

Tableau 1. Type de précisions requises pour les données géographiques

Validité des informations

16. Le *data paper* doit rendre compte du processus de validation de la géodatabase qui garantit sa qualité (validation technique, analyses de sensibilité et de robustesse, comparaisons avec d'autres référentiels de données, etc.). Il faut mentionner le niveau d'adéquation entre les résultats de ce processus de validation et les objectifs initiaux de la construction de la géodatabase. Il est nécessaire d'indiquer très clairement les points forts et les limites du jeu de données publié par rapport à d'autres matériaux existants, le cas échéant. Dans le cas particulier des données géographiques, la comparaison des données avec d'autres jeux existant sur le même espace d'étude, ou avec des données similaires sur d'autres terrains, est encouragée.

Valeur ajoutée

17. La valeur ajoutée de cette base de données à la connaissance en géographie et plus largement en sciences sociales est un des critères déterminants pour sa publication, il est demandé aux auteurs de bien mettre en valeur ce point qui représente l'apport principal de leur contribution scientifique. Un autre apport important consiste dans le mode d'emploi de la base de données, qui donne un avantage spécifique aux futurs utilisateurs potentiels. Dans le cas particulier des données géographiques, une évaluation de la possibilité d'étendre l'envergure géographique des données (à d'autres pays, régions, municipalités, etc.) est appréciée.

Adaptation de l'organisation d'une revue scientifique pour inclure des *data papers*

18. En plus de produire les recommandations exposées ci-dessus pour ses auteurs et ses évaluateurs, la revue a dû procéder en 2017 à un certain nombre d'innovations pour implémenter et consolider la nouvelle rubrique des *Data Papers*. Cela constitue une nouvelle étape dans la constante progression de *Cybergeo* sur le chemin de la science ouverte, vers la reproductibilité des savoirs publiés. L'expérience antérieure de la création en 2014 d'une rubrique GeOpenMod, dédiée à la publication de modèles géographiques (mettant à disposition les codes, dockers et métadonnées), a facilité cette nouvelle adaptation.

Un nouveau comité de lecture

19. Dans l'objectif de mettre en responsabilité des personnes ayant les compétences nécessaires pour l'évaluation, la revue a constitué un comité de lecture spécifique à cette rubrique. Les compétences requises portent sur la connaissance des sources de données et des bases de données similaires préexistantes, sur les bonnes pratiques en termes de méthodes de construction et de validation des données géographiques et des moyens favorisant leur réutilisation. Une douzaine de personnes, géographes, statisticiens, informaticiens, géomaticiens ou ingénieurs, ont accepté de faire partie de ce comité et rendent des avis bien informés sur ces articles. Selon le thème des articles, il est aussi fait appel à des évaluateurs extérieurs spécialisés. Mais la revue a tenu à donner une limite à ces expertises, en signalant aux auteurs et aux lecteurs qu'elle ne procède pas à une validation des données elles-mêmes (car elle n'en a pas les moyens), mais seulement à la vérification que les données peuvent être réutilisées pour reproduire les résultats annoncés par les auteurs de l'article (cette vérification exige déjà souvent un travail important de la part des évaluateurs). En cela, la philosophie de publication des données de *Cybergeo* se situe en position intermédiaire dans la distinction faite par Callaghan *et al.* (2012) entre *published data* – dont le seul critère est d'être accessibles – et *Published data*, dont les qualités sont d'être accessibles de manière pérenne, documentées et évaluées par les pairs. *Cybergeo* réserve l'évaluation par les pairs aux articles, et non aux données elles-mêmes.

20. L'évaluation des *data papers* (à l'instar des *model papers* de la rubrique GeOpenMod), mobilise donc généralement (au moins) un relecteur du comité de lecture dédié, dont l'attention se porte principalement sur la qualité de la production et de la transformation des données, et (au moins) un relecteur externe au comité de lecture des *data papers*, mobilisé pour son expertise thématique, et dont le rôle est d'évaluer l'intérêt du résultat et la justification thématique des choix opérés.

Pérennité et ouverture des entrepôts de données

21. La revue admet que les auteurs aient déposé leurs données dans d'autres entrepôts que celui propre à la revue, mais insiste pour que ces institutions garantissent la pérennité des données et soient ouverts à leur réutilisation. On demande aux auteurs de privilégier les dépôts institutionnels (par ex. Figshare, Zenodo, Nakala…) et d'exiger l'attribution d'un DOI. Après avoir exploré un certain nombre de sites possibles, la revue privilégie le Dataverse de l'université d'Harvard où elle dispose désormais d'une collection en son nom pour accueillir les données de ses *data papers*. En construisant un partenariat avec ce centre, elle a apporté en échange une innovation dans le protocole des dépôts afin de résoudre la question de l'anonymisation pour l'évaluation. La revue dispose également de collections sur Zenodo (Europe) et Nakala (France) qui ont été créées à la suite de dépôts sur ces entrepôts de jeux de données par plusieurs auteurs. Le lancement en 2022 de la plateforme Recherche Data Gouv ayant pour objectif de fédérer les données de la recherche française et de permettre à la communauté scientifique française de publier les données de recherche issues de la recherche publique ouvre de nouvelles perspectives. À terme, lorsque la plateforme sera ouverte à la géographie, les jeux de données des différentes collections de *Cybergeo* devraient y être accessibles. Par ailleurs, nous étudions la possibilité d'ouvrir une collection dans le Dataverse de Recherche Data Gouv si les conditions d'anonymisation au moment de l'évaluation peuvent être respectées.

Anonymisation des données pour l'évaluation

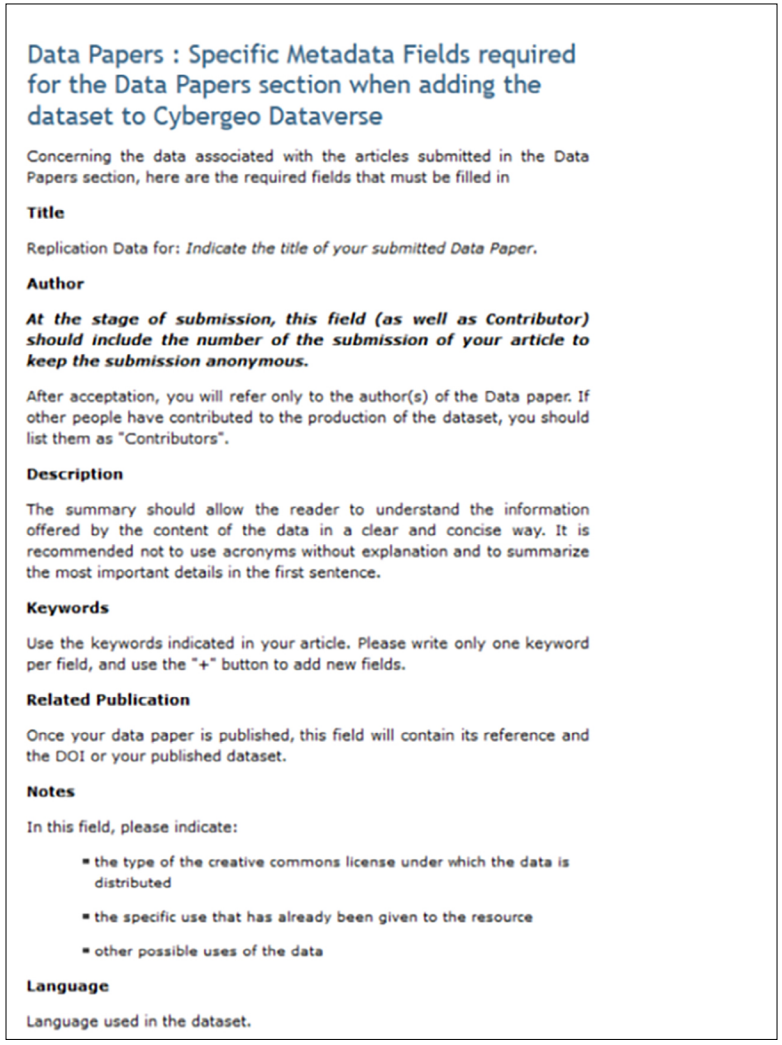
22. En effet, pour pratiquer une évaluation en double aveugle, analogue à celle qui se pratique pour les autres articles, la revue a négocié avec Harvard Dataverse l'introduction d'une fonctionnalité qui permet aux évaluateurs des articles, pour la collection de *Cybergeo*, de télécharger les données sans connaître les identités des auteurs.

Introduction de nouvelles métadonnées pour le géoréférencement

23. Contrairement à la pratique très libérale et peu normée de présentation des articles en SHS, la publication de *data papers* a conduit la revue à proposer des normes très précises qui contraignent les auteurs à apporter les précisions nécessaires à la compréhension et à la réutilisation de leurs données (Tableau 2). Ces métadonnées incluent notamment un champ géographique permettant de faire des recherches d'articles à partir de l'espace d'étude concerné (au niveau pays et continents[5]). Dans l'ensemble, les auteurs comprennent la nécessité de remplir correctement les paragraphes requis pour permettre la réutilisation de leurs données et bien valoriser leur apport. Deux ou trois allers-retours entre la soumission de l'article et les demandes des évaluateurs sont cependant nécessaires en moyenne pour que

5. https://journals.openedition.org/cybergeo/28086

toutes les précisions demandées figurent dans la version finalement publiée.



Data Papers : Specific Metadata Fields required for the Data Papers section when adding the dataset to Cybergeo Dataverse

Concerning the data associated with the articles submitted in the Data Papers section, here are the required fields that must be filled in

Title

Replication Data for: *Indicate the title of your submitted Data Paper.*

Author

At the stage of submission, this field (as well as Contributor) should include the number of the submission of your article to keep the submission anonymous.

After acceptation, you will refer only to the author(s) of the Data paper. If other people have contributed to the production of the dataset, you should list them as "Contributors".

Description

The summary should allow the reader to understand the information offered by the content of the data in a clear and concise way. It is recommended not to use acronyms without explanation and to summarize the most important details in the first sentence.

Keywords

Use the keywords indicated in your article. Please write only one keyword per field, and use the "+" button to add new fields.

Related Publication

Once your data paper is published, this field will contain its reference and the DOI or your published dataset.

Notes

In this field, please indicate:

- the type of the creative commons license under which the data is distributed
- the specific use that has already been given to the resource
- other possible uses of the data

Language

Language used in the dataset.

Tableau 2 : Un template exigeant pour la rédaction des *data papers* en géographie

https://journals.openedition.org/cybergeo/23412#tocto2n5

Conclusion

24. Les résultats de cette innovation peuvent sembler modestes, avec une dizaine d'articles publiés depuis 2018 dans la rubrique des *Data Papers*. Mais d'après les statistiques d'OpenEdition, ils ont reçu chacun en 2021 plus de 500 vues, et le plus consulté, publié en 2020, l'a été près de 3 000 fois.

25. Il est encore trop tôt pour évaluer un impact en termes de citations (26 en 2021) et surtout de réutilisations des données. Cependant, la diversité des thèmes de ces bases de données géographiques (évolution de la population des agglomérations en Chine et en Europe ; délimitations de grandes régions urbaines mondiales utiles aux travaux sur la mondialisation des entreprises et sur les réseaux de recherche ; diffusion spatiale de l'information d'après une base historique des journaux néerlandais ; déterminants politiques et naturels du classement d'AOC viticoles dans une région française ; variations et évolution des prix immobiliers en Île-de-France et dans des villes européennes comparé au pouvoir d'achat des ménages, etc.) témoigne tout à la fois de l'éclectisme des centres d'intérêt, de l'utilité de la rubrique et d'une appétence pour le processus de publication proposé.

26. Quelques difficultés ont été rencontrées dans les premiers temps de ce type de publication. La première tenait au dédoublement des articles, proposés à la revue en même temps sous des titres voisins, l'un dans une rubrique « classique » et l'autre comme *data paper*. Les

auteurs hésitaient en effet entre le souhait de partager leurs données et l'envie légitime d'en publier une première valorisation des résultats. La revue a parfois accepté de publier les deux versions, lorsqu'elles étaient suffisamment autonomes et peu redondantes. Parfois, les auteurs ont été convaincus de réintégrer la description de leurs résultats dans le *data paper*, pour lequel ces résultats apportaient une bonne démonstration de l'utilité et de la pertinence de ces données. Une difficulté qui a été un peu plus longue à surmonter, mais s'est avérée génératrice d'une solution de qualité est liée au problème d'anonymisation du dépôt des données pour permettre l'évaluation en double aveugle, qui a conduit à faire bénéficier la revue de l'expérience d'un des plus grands centres internationaux de dépôt de données scientifiques en SHS. Enfin, il est parfois difficile de trouver des évaluateurs ayant à la fois une connaissance précise des sources d'information et des méthodes de traitement des données, exigée pour ce type d'article, et une foi suffisante dans l'avenir de la science ouverte pour accepter de consacrer le temps nécessaire à ces évaluations un peu plus exigeantes en termes de précision et d'analyse.

27. La revue *Cybergeo* a été dès 2017 la première des revues internationales de géographie à proposer la publication de *data papers*. Elle a été suivie en 2021 par la prestigieuse revue britannique *Environment and Planning B: Urban Analytics and City Science* (Arribas-Bel *et al.*, 2021), qui regroupe d'ailleurs, au sujet des données urbaines sous l'appellation de Urban Data/Code, l'équivalent des rubriques *GeOpenMod* et *Data Papers* de *Cybergeo*. Cette ouverture vers la science ouverte laisse augurer d'une meilleure prise de conscience de la part des éditeurs britanniques de l'aspiration des chercheurs au partage de la science.

28. En effet, la banalisation attendue des articles du type *data paper* doit aussi aider à ce que l'énorme travail de création et d'adaptation de données soit de mieux en mieux pris en compte dans les évaluations des travaux des laboratoires et des chercheurs, ce qui devrait contribuer à lever les hésitations des auteurs, et des lecteurs, en géographie comme ailleurs.

La FAIRisation des données

Alain Rivet

Introduction

1. La gestion rigoureuse et cohérente des données de la recherche constitue aujourd'hui un enjeu de taille pour la production de nouvelles connaissances scientifiques. Il est important, tout d'abord, de noter qu'au sein d'un établissement public, les données et documents produits dans l'exercice des activités du personnel constituent des données et documents relevant du Code du Patrimoine et appartiennent de fait à l'établissement.

2. Guidés par le « Plan National pour la science ouverte » (ministère de l'Enseignement supérieur de la Recherche et de l'Innovation 2021), les différents organismes de recherche participent à la réflexion sur cette question afin de mettre à disposition des outils, méthodes et infrastructures répondant aux besoins des communautés scientifiques en matière de gestion et de partage des données scientifiques. Ce plan fournit les orientations en matière de gestion *FAIR* des données (*Findable, Accessible, Interoperable, Reusable*) soit Faciles à trouver, Accessibles, Interopérables et Réutilisables.

3. Nous nous proposons, dans ce chapitre, d'apporter une réponse pratique à cette nécessité de participer au mouvement de la science ouverte face à l'explosion des données et l'organisation de nos unités de recherche. Pour ce faire, nous nous sommes appuyés sur les travaux du réseau Qualité en Recherche et du groupe Atelier-Données de la Mission pour les initiatives transverses et interdisciplinaires (MITI) du CNRS, qui sont à l'origine de deux guides, respectivement le guide « Traçabilité des activités de recherche et gestion des connaissances – Guide pratique de mise en place » (Rivet, Bachèlerie, Denis-Meyere et Tisserand, 2018) et le « Guide de bonnes pratiques sur la gestion des données de la recherche » (Hadrossek, Janik, Libes, Louvet, Quidoz, Rivet et Romier, 2021).

4. Nous aborderons, dans un premier temps, la problématique de la croissance des données de la recherche, les enjeux du partage de ces données, la qualité et la traçabilité des activités de recherche avant d'aborder, dans un deuxième temps, les moyens de parvenir à une gestion *FAIR* des données à travers la mise en place de bonnes pratiques de gestion via la rédaction d'un plan de gestion de données (PGD) ou *Data Management Plan* (DMP), d'un *data paper*, l'utilisation de cahiers de laboratoire électroniques, l'archivage et le dépôt dans des entrepôts de données.

Les enjeux de la gestion des données

Gérer l'explosion des données de la recherche

5. Il convient, au préalable, de préciser ce que l'on entend par « données de la recherche ». La définition des données de la recherche de l'OCDE (Organisation de coopération et de développement économiques, 2007) est la suivante : « Enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche » ; « ce terme ne s'applique pas aux éléments suivants : carnets de laboratoire, analyses préliminaires et projets de documents scientifiques, programmes de travaux futurs, examens par les pairs, communications personnelles avec des collègues et objets matériels (par exemple, les échantillons de laboratoire, les souches bactériennes et les animaux de laboratoire tels que les souris). »

6. Cette définition qui exclut, de fait, les données administratives, supports divers, cahiers de laboratoire, etc., nous apparaît trop restrictive et nous intégrerons, dans notre approche, l'ensemble des informations d'une unité de recherche, qu'il s'agisse d'échantillons, d'enquêtes, de données administratives et scientifiques tant manuscrites (cahiers de laboratoires, etc.) que numériques (figure 1).

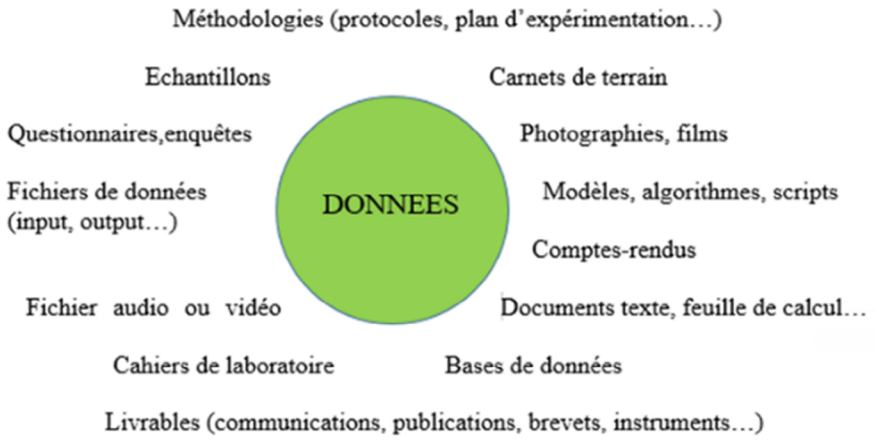


Figure 1. La diversité des données de la recherche.

7. Dans le même temps, dans un monde aujourd'hui largement numérique, nous sommes entrés dans l'ère du *Big Data* et devons faire face à une très forte croissance des données produites dans le monde.

8. Selon les dernières estimations (figure 2), le volume de données numériques créées ou répliquées à l'échelle mondiale a été multiplié par plus de trente au cours de la dernière décennie, avec une prévision de 180 zettaoctets (un zettaoctet = 10^{21} octets) à l'horizon 2025.

9. La science n'est pas en reste et la recherche engendre également une croissance extrêmement forte de ses données, quels que soient les secteurs d'activité, du fait de la dématérialisation de nombreuses activités administratives, la production massive de données par les équipements scientifiques, etc. Les expériences du *Large Hadron Collider* (LHC) du CERN produisent ainsi

de l'ordre de 90 pétaoctets (1 pétaoctet = 10^{15} octets) de données par an.

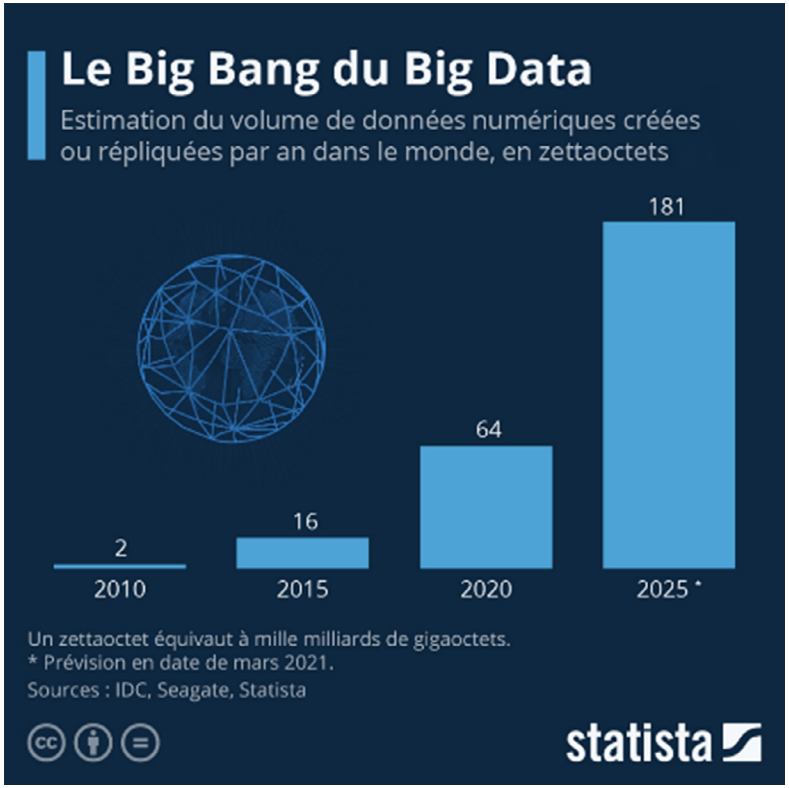


Figure 2. Le *Big Bang* du *Big Data*.
Crédit : Gaudiaut, 2021.

10. Ces données sont souvent riches en informations, car obtenues suivant une démarche scientifique précise. Certaines sont uniques et donc irremplaçables, comme les premiers relevés satellites de la banquise des années 1960, ou possèdent une forte valeur patrimoniale, comme les données archéologiques de la cité antique de Palmyre en Syrie ou la cartographie 3D de la charpente de Notre-Dame réalisée en 2014. Les données numériques représentent donc un enjeu majeur de la recherche, certains les comparant à « l'or noir du XXI^{e} siècle » et il est essentiel de réfléchir à leur devenir et d'assurer leur pérennité.

Répondre aux enjeux du partage des données

11. Les enjeux du partage des données sont multiples. Scientifiques tout d'abord pour permettre l'exploitation des résultats de recherche antérieurs, économiques de façon à éviter la duplication des efforts de recherche, mais aussi sociétaux en termes de transparence du processus scientifique.

12. Le grand collisionneur électron-positron (LEP) du CERN s'est arrêté en 2000 pour permettre la construction de son successeur, le LHC. Près de 300 publications ont été produites après son arrêt ; c'est un phénomène assez récurrent en recherche, les publications continuent longtemps après la fin des expérimentations avec l'émergence de nouvelles idées, de nouvelles théories et la confrontation avec d'autres études.

13. Par ailleurs, il est généralement considéré que seulement 10 % des données de la recherche sont utilisées par les publications scientifiques, les données publiées dans les articles scientifiques représentant seulement la « partie émergée de l'iceberg ». Selon plusieurs études, les données

se perdraient à un rythme inquiétant ; une étude souligne que deux années après la publication d'un article, les chances d'accéder aux données scientifiques de l'étude chutent de près de 17 % par an (Vines T. *et al.*, 2014).

14. Cela signifie qu'un chercheur produit beaucoup plus de données que celles qui, *stricto sensu*, sont nécessaires pour valider les résultats de la recherche, des données qui pourraient être utilisées par d'autres chercheurs dans le cadre d'un nouveau projet de recherche. C'est d'ailleurs cet usage que le mouvement d'ouverture des données souhaite faciliter. Dès lors, il devient indispensable de mettre en place des mécanismes pour en assurer leur réutilisation future.

Disposer de données de qualité

15. Dès lors que l'on est convaincu de la nécessité de conserver et partager les connaissances, encore faut-il que ces données partagées soient des données de « qualité ». En effet, en 2021 à l'occasion du printemps de la plateforme data.gouv.fr, la question de la qualité des données s'est posée, partant du postulat que l'ouverture des jeux de données n'entraînait pas directement leur réutilisation.

16. Ainsi, « l'analyse de l'enquête auprès des usagers (905 répondants de juin à septembre 2020) pointe une véritable attente des utilisateurs de la plateforme sur la qualité des données. Les répondants remontent des problèmes de mise à jour avec des jeux de données souvent obsolètes, une documentation insuffisante ou inexacte quand elle existe, la multiplicité de jeux de données, etc. En somme, la qualité n'est pas suffisamment au rendez-vous » (Data.gouv.fr, 2021).

17. La qualité des données peut en fait être dégradée à deux niveaux, au niveau de la donnée elle-même avec des valeurs anormales, obsolètes, des informations non représentatives que l'on associe classiquement aux notions de reproductibilité et de fiabilité des résultats, mais aussi de la description de ces données que sont les métadonnées (données sur les données).

18. Le premier niveau pose la question de la qualité intrinsèque des données, directement associée au déroulement de la recherche. En effet, par nature, la recherche n'est pas répétitive, mais riche en incertitudes. La confiance dans la qualité d'une recherche consiste à établir et vérifier que les différentes étapes d'une étude peuvent être répétées en obtenant le même résultat par des chercheurs différents à des moments différents. Il est ainsi essentiel de s'assurer que l'ensemble des activités qui conduit à l'obtention d'une donnée est maîtrisé, cela est par exemple le cas de la chaîne fonctionnelle d'une analyse (des pipettes, balances jusqu'aux équipements d'analyse) ou du respect d'un protocole d'enquête (échantillon représentatif, etc.).

19. Le second point concerne l'environnement de la donnée c'est-à-dire le niveau de qualité du jeu de données qui va être partagé et qui peut être évalué par la structure des données (structure et format du fichier), le respect de

standards et référentiels reconnus et la présence d'informations complémentaires (documentation, etc.).

Garantir l'intégrité scientifique

20. Une gestion *FAIR* des données de la recherche contribue également à assurer l'intégrité scientifique et prévenir la fraude scientifique. En effet, on observe actuellement que l'« évaluation par les pairs » sur laquelle est fondé le principe fondamental de publication scientifique semble montrer ses limites (Pigenet et Ben Ytzhak, 2014).

21. L'incapacité à apporter la preuve d'un élément clé d'une publication apparaît en effet comme un risque majeur dans le contexte actuel de la recherche. En effet, la pression qui s'exerce sur les chercheurs (*publish or perish*) et l'augmentation croissante du personnel temporaire dans les structures de recherche du fait des nouveaux modes de financement de la recherche (ANR, Europe, etc.) se conjuguent avec une médiatisation de plus en plus forte des recherches scientifiques du fait de leur impact sociétal.

22. Publiée en janvier 2015, la charte nationale de déontologie des métiers de la recherche, signée par de très nombreux établissements scientifiques (figure 3), s'adresse à l'ensemble des personnels d'un établissement ou d'un organisme qui contribuent à l'activité de recherche avec l'objectif « d'expliciter les critères d'une démarche scientifique rigoureuse et intègre, applicable notamment dans le cadre de tous les partenariats nationaux et internationaux » (Établissements de recherche et d'enseignement supérieur français, 2015).

Figure 3. Signataire de la charte française de déontologie des métiers de la recherche.

Crédit : Office français de l'intégrité scientifique (https://www.ofis-france.fr/la-charte-francaise-de-deontologie-des-metiers-de-la-recherche/).

23. Deux extraits de cette charte concernent directement notre propos :
    - « la description détaillée du protocole de recherche dans le cadre des cahiers de laboratoire, ou de tout autre support, doit permettre la traçabilité des travaux expérimentaux » ;
    - « tous les résultats bruts (qui appartiennent à l'institution) ainsi que l'analyse des résultats doivent être conservés de façon à permettre leur vérification ».

24. Plus récemment, citons le décret du 3 décembre 2021 (Légifrance, 2021) relatif au respect des exigences de l'intégrité scientifique par les établissements publics, dont l'article 6 précise : « Les établissements publics et fondations reconnues d'utilité publique mentionnés au troisième alinéa de l'article L. 211-2 du code de la recherche définissent une politique de conservation, de communication et de réutilisation des résultats bruts des travaux scientifiques menés en son sein. À cet effet, ils veillent à la mise en œuvre par leur personnel de plans de gestion de données et contribuent aux infrastructures qui permettent la conservation, la communication et la réutilisation des données et des codes sources. »

Les bonnes pratiques de gestion des données

Appliquer les principes FAIR

25. Dès lors, comment mettre en place une gestion *FAIR* des données dont les principes ont été publiés en 2016 (Wilkinson Mark *et al.*, 2016). L'*open science* introduit cette notion de *FAIR* data (figure 4) dont les principes permettent de guider les stratégies de gestion des données et d'aider tous les acteurs qui œuvrent à les produire, à en contrôler la qualité, à les traiter, à les analyser, à assurer leur publication et leur diffusion sur des plateformes de partage ou d'archivage.

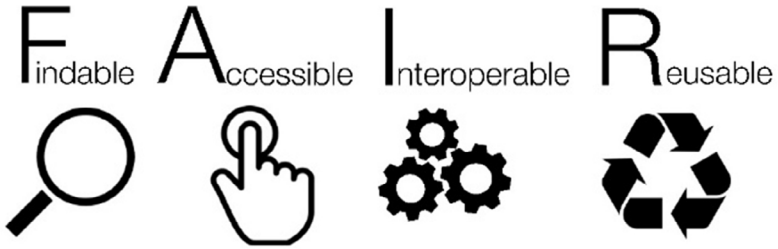


Figure 4. Les 4 principes du *Fair data*.
Crédit : Sangya Pundir [CC BY-SA 4.0]).

26. Toutefois, s'il y a une volonté forte en faveur du partage et de la réutilisation des données, les principes *FAIR* n'impliquent pas l'ouverture systématique des données, le principe de base « aussi ouvert que possible, aussi fermé que nécessaire » restant en vigueur. Cela signifie que la

propriété intellectuelle, dans le cadre d'une recherche partenariale, prime sur l'ouverture des données, un taux de financement de 50 % sur des fonds publics conditionnant cette ouverture.

27. Divers principes directeurs dérivent de chacune des lettres du mot *FAIR* :

- *Findable* : les données doivent être faciles à trouver ou à retrouver. À cet effet, celles-ci doivent disposer d'un identificateur unique et pérenne et être décrites au moyen de métadonnées. Données et métadonnées doivent être enregistrées et indexées de façon à permettre leur recherche et sont préservées dans le temps.
- *Accessible* : l'utilisateur, une fois les données trouvées, doit en connaître les modalités d'accès. Les données et leurs métadonnées doivent être accessibles par leur identifiant via un protocole de communication standardisé (ouvert, libre et d'usage universel). Les conditions d'accès des licences (Creative Commons par exemple) doivent être connues et visibles, tout comme les procédures d'authentification et d'autorisation, lorsque cela s'avère nécessaire.
- *Interoperable* : l'interopérabilité est « la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs, et ce, sans restriction d'accès ou de mise en œuvre » (AFUL, 2015). Dans son application aux données, l'interopérabilité nécessite d'utiliser des protocoles d'accès et des formats de données ouverts, normés ou standardisés, à la fois au niveau des formats de fichiers, mais également à travers les outils informatiques qui serviront à échanger, diffuser et lire les données.
- *Reusable* : la réutilisation (libre ou conditionnelle) doit être facilitée par l'utilisation de standards communs grâce à des bases de données rassemblant des données claires, vérifiées et bien décrites. Les données doivent être richement décrites, par une pluralité d'attributs précis et pertinents, incluant des détails sur leur provenance. Elles doivent disposer d'une licence d'utilisation claire et accessible.

28. Les *data papers* contribuent notamment aux principes F (*findable*) et R (*reusable*), par une description riche des données, de leur intérêt et des conditions de leur réutilisation.

Se baser sur le cycle de vie des données

29. Une gestion *FAIR* des données est un processus complexe qui nécessite plusieurs étapes avant d'aboutir à la publication et au partage des données. Pour formaliser ces différentes étapes, il est possible de s'appuyer sur le cycle de vie des données, qui est un cercle vertueux correspondant aux différentes phases d'un projet scientifique et dont une version a été élaborée au sein de l'Atelier Données (figure 5).

30. Pour chacune des étapes de ce cycle, il convient de mettre en place diverses actions permettant d'arriver à une gestion *FAIR* des données, très largement développées dans le guide de bonnes pratiques sur la gestion des données de la recherche.

Figure 5. Le cycle de vie des données de la recherche.
Crédit : Hadrossek, Janik, Libes, Louvet, Quidoz, Rivet et Romier, 2021.

31. Les deux premières phases 1 – Imaginer – et 2 – Concevoir et planifier –, sont les étapes préparatoires d'un projet où l'on se préoccupe de disposer de toutes les informations nécessaires à la bonne gestion des données et du projet. C'est l'étape où l'on réfléchit au plan de gestion de données, où l'on prépare les espaces de stockage et où l'on met en place les outils de gestion de projet. Cette partie, très générique, a pour objectif de conduire le personnel à s'interroger sur ses besoins et les moyens dont il dispose pour trouver des solutions adaptées. C'est à ce niveau que débute la mise en place du PGD/DMP.

32. Différentes initiatives institutionnelles sont développées au sein de nos établissements pour accompagner la politique des données de la recherche. Citons quelques exemples en région Auvergne-Rhône-Alpes :

1. le projet DATACC[1], dispositif d'accompagnement sur les données de recherche en physique et chimie porté par les bibliothèques de Lyon 1 et de Grenoble Alpes Université ;
2. la Cellule Data Grenoble Alpes, structure opérationnelle proposant des services de proximité pour l'accompagnement des communautés scientifiques du site de Grenoble Alpes sur tout le cycle de vie de la donnée[2].

33. Sur un plan national, des outils sont également mis à la disposition de la communauté scientifique :

1. Le projet DoRANum[3] (Données de la Recherche : Apprentissage NUMérique) fournit des ressources pour accompagner la communauté scientifique dans la gestion et le partage de leurs données ;
2. CatOPIDoR[4] est un wiki des services dédiés aux données de la recherche.

34. L'étape 3 – Collecter – du cycle de vie concerne les aspects d'acquisition et de collecte des données (capteurs, instruments, sondages, enquêtes, etc.) ainsi que la constitution des jeux de données (*dataset*) avec leurs métadonnées associées. La description de ces jeux de données nécessite

1. https://www.datacc.org/
2. https://scienceouverte.univ-grenoble-alpes.fr/
3. http://doranum.fr/
4. https://cat.opidor.fr/

d'utiliser, dans la mesure du possible, des référentiels de vocabulaires contrôlés (thésaurus), standardisés et appropriés au domaine étudié. La collecte des données se fait dans le respect des règles de traitement spécifiques des données personnelles. Un éclairage sur les environnements de stockage est apporté.

35. L'étape 4 – Traiter – témoigne du prétraitement des données brutes acquises et collectées précédemment. La connaissance et la maîtrise des formats et standards sont importantes. Il s'agit souvent de vérifier, regrouper et qualifier les données pertinentes parmi celles qui ont été collectées, puis de les reformater dans des formats standards interopérables et les préparer pour leur analyse ultérieure.

36. L'étape 5 – Analyser – consiste généralement à extraire l'information des données récoltées par l'utilisation de nombreux outils et techniques (calcul intensif, traitement statistique, machine learning, visualisation, etc.). Cette étape impose que ces données soient exploitables, c'est-à-dire bien organisées, dans des formats adaptés à l'analyse envisagée, en particulier à des fins de traitements automatisés.

37. L'étape 6 – Préserver et archiver – rend compte de l'importance de préserver et archiver les données sur le long terme. On s'attache à réfléchir aux données pertinentes à préserver et à étudier les solutions à mettre en œuvre dans son environnement.

38. L'étape 7 – Publier et diffuser – est la phase finale du cycle de vie permettant de diffuser les données à travers des catalogues de données, des *thesaurus* de mots clés, des identifiants pérennes, des entrepôts de données et des *data papers*.

39. L'article de données ou *data paper* permet de rendre plus visibles tous les jeux de données d'une recherche présentant un potentiel de réutilisation important. Alors qu'une publication scientifique analyse et interprète les données scientifiques, un *data paper* décrit précisément un ou plusieurs jeux de données de façon à en faciliter la compréhension et l'éventuelle réutilisation. Il décrit les modalités d'acquisition des données, leurs conditions et droits d'utilisation et contient la description de toutes les métadonnées associées de façon à en garantir la qualité.

40. Cette dernière étape d'un projet de recherche représente en quelque sorte la finalité de toute politique de gestion de données, puisqu'elle vise à publier et diffuser les données de manière à ce qu'elles soient accessibles et réutilisables selon des formats et des processus interopérables, via des identifiants pérennes lors du dépôt dans des entrepôts de données.

41. Parmi ces différentes étapes du cycle de vie, nous allons faire un focus sur quelques items essentiels pour une gestion *FAIR* des données en abordant les plans de gestion de données, les outils de traçabilité comme les cahiers de laboratoire électroniques, l'organisation des

données et leur conservation (sauvegarde, archivage, entrepôts de données).

Rédiger un plan de gestion de données (PGD)

42. Le PGD/DMP est un document formalisé permettant de recenser les informations nécessaires à la création, la mise à disposition, la maintenance, la conservation et la protection des données d'un projet de recherche. Cela nécessite de renseigner les différents items qui constituent la trame d'un PGD (Deboin, 2018) :
    - Informations administratives
    - Collection de données
    - Documentation et métadonnées
    - Éthique, cadre légal
    - Stockage, sauvegarde, sécurité
    - Sélection et conservation
    - Partage des données
    - Responsabilités et moyens

43. L'outil DMP OPIDoR[5] est une application web permettant de réaliser un plan de gestion de données en ligne à partir des modèles proposés (ANR, Europe, etc.).

Disposer d'un outil de traçabilité : le cahier de laboratoire

44. L'ensemble des données produites dans une activité de recherche doit être répertorié et enregistré dans l'objectif d'assurer une traçabilité des différentes activités réalisées. Nous disposons, pour ce faire, de supports tels que les cahiers de laboratoire.

45. Le cahier de laboratoire constitue un véritable outil scientifique, et ce, dès le commencement d'un projet. Il permet d'assurer l'intégrité scientifique tout en répondant aux obligations légales et contractuelles, en apportant la preuve de l'invention et de ses inventeurs.

46. En 2007, un cahier de laboratoire national[6] a été mis en place en lien avec les préconisations du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche (MENESR) sur la propriété intellectuelle de façon, d'une part, à renforcer l'esprit d'appartenance à la recherche publique française et d'autre part, à donner une image d'excellence et de confiance de la recherche française.

47. Toutefois, face à l'explosion des données numériques, l'« outil papier » (figure 6) apparaît de moins en moins adapté aux pratiques de la recherche. Les apports du numérique sont en effet multiples à travers le partage d'informations, l'insertion de fichiers de données,

5. https://opidor.fr/

6. https://www.curie.asso.fr/-Cahier-de-laboratoire-national-.html

une recherche d'informations facilitée, une datation des expériences par l'horodatage tout en facilitant le travail collaboratif.

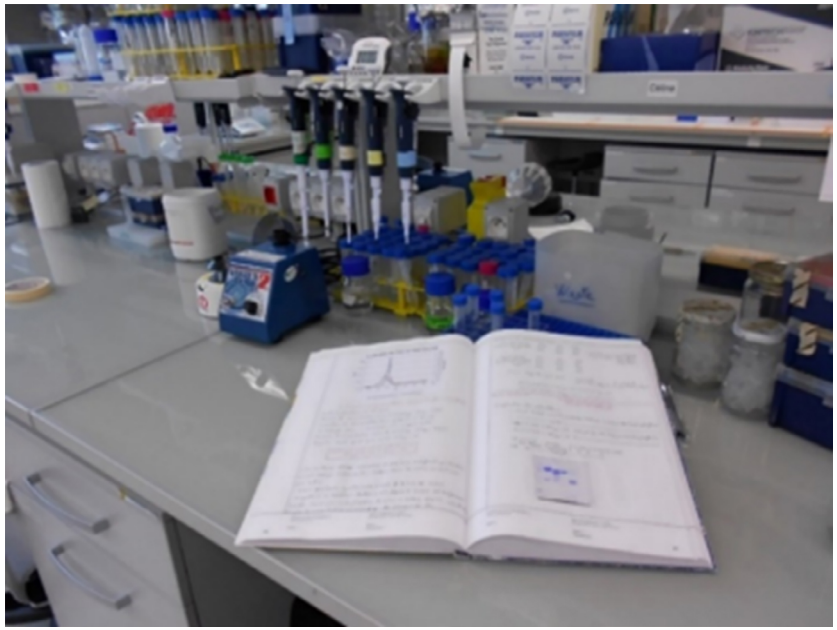


Figure 6. Utilisation d'un cahier de laboratoire papier

48. Le cahier de laboratoire électronique (CLE) ou *Electronic Lab Notebook* (ELN) va également permettre de disposer de toutes les informations utiles à la description des données (libellés des paramètres, unités de mesure, localisation, étiquettes, propriétaires, etc.) ainsi que sur les dispositifs d'acquisition (équipements, capteurs de mesures, modèles numériques, etc.) qui vont ainsi constituer des métadonnées.

49. La problématique des cahiers de laboratoire électroniques s'est intensifiée ces dernières années. En 2021, le groupe de travail « Cahier de laboratoire électronique » du comité pour la science ouverte (CoSO) a rédigé un rapport destiné à présenter une vision partagée sur la définition, le cadrage, les usages et le périmètre fonctionnel du CLE, qui doit pouvoir s'intégrer dans les environnements informatiques et institutionnels existants. Le rapport fournit un ensemble de recommandations sur les critères de choix d'un tel outil et livre une liste comparative de logiciels (Comité pour la science ouverte 2021).

50. Dans le domaine des SHS, le CLE dans sa version mobile (tablette, téléphone) peut remplacer efficacement les supports papier pour la prise de notes sur le terrain, avec, par exemple, un mode « hors ligne » assurant une synchronisation différée des données. De plus, avec le CLE, la discipline profitera largement des aspects de sécurisation des données, d'indexation avec une recherche d'informations facilitée sans compter les possibilités d'interfaçage avec d'autres outils numériques.

51. Dans le cycle de vie des données (figure 5), le CLE intervient pendant toute la phase de réalisation interne du projet de recherche, de l'étape 2 (Concevoir – Planifier) à l'étape 5 (Réaliser – Analyser). Il apparaît, de ce fait, une réponse pertinente aux enjeux de reproductibilité des recherches et de gestion des données attenantes.

Mettre en place une organisation des données

52. Le cahier de laboratoire électronique, s'il permet d'intégrer divers documents (textes, images, etc.) ne permet pas à l'heure actuelle le stockage, en son sein, de grosses quantités de données que constituent souvent les résultats des expériences. Ce besoin peut s'exprimer en pétaoctets et se heurte techniquement à l'heure actuelle à des problématiques réseau. De ce fait, il conviendra de définir et mettre en place une organisation des données brutes non intégrées, mais référencées dans le CLE.

53. La réutilisation de ces données se fait en ajoutant de l'intelligibilité au stockage des données. À cet effet, il conviendra de définir un plan de classement des dossiers sur les espaces de stockage référencés et d'identifier efficacement les fichiers numériques en fixant des règles de nommage.

Définir un plan de classement

54. Il est important de disposer d'un plan de classement des dossiers numériques pour améliorer la recherche ultérieure de fichiers. Ce plan de classement n'est pas figé et peut évoluer selon les données à enregistrer. Il devra disposer d'une hiérarchie de répertoires et de dossiers (limités à cinq ou six niveaux hiérarchiques) avec des intitulés intelligibles par tous allant du général au particulier. Il est par exemple possible de choisir une organisation thématique aux niveaux supérieurs, puis alphabétique et/ou chronologique pour les dossiers des niveaux inférieurs.

Fixer des règles de nommage

55. Il est capital d'élaborer et de mettre en place des règles communes de nommage des fichiers numériques pour faciliter et pérenniser l'accès à l'information ainsi que pour optimiser le partage et le tri des documents. Chaque unité doit définir la méthodologie adoptée, la formaliser et la communiquer à l'ensemble du personnel. Pour ce faire, les intitulés des fichiers doivent être uniques et succincts (ne pas excéder 31 caractères, extension comprise). Ils peuvent être caractérisés *a minima* par une date, un sujet et un type de document. Il convient d'éviter accents, caractères spéciaux, mots vides (le, la, les, un, une, des, et, etc.) et les dénominations vagues (« divers », « autres », etc.). Les espaces doivent être remplacés par des « _ ».

Conserver les données

56. Stocker, sauvegarder, archiver sont des phases essentielles d'une gestion rigoureuse des données. Toutefois, il n'est pas toujours aisé de faire la distinction entre ces notions.

57. Le stockage sert des usages de partage de fichiers et de documents. C'est généralement l'étape première qui consiste à déposer les données sur un support numérique pour les rendre accessibles. À ce stade, la

donnée n'est ni sauvegardée ni sécurisée et reste de fait fragile. En effet, une étude portant sur 25 000 ordinateurs a montré que 20 % des disques durs cessent de fonctionner après 4 ans d'utilisation.

58. La sauvegarde consiste à dupliquer des données à l'identique pour pouvoir les restaurer en cas de dommage ou de perte. Une sauvegarde est régulière et les données stockées sont régulièrement modifiées. Cette étape de sauvegarde doit s'accompagner d'une réelle politique de sauvegarde qui détermine, par exemple, la fréquence de sauvegarde en fonction de la criticité et de la sensibilité des données. À ce stade, la préservation de l'intelligibilité des données n'est pas un élément pris en compte.

59. L'archivage a pour finalité de préserver les données anciennes sur une longue durée à des fins de référence. L'archivage consiste à rendre accessible en lecture des données immuables (archives de documents administratifs, données de mesures expérimentales, résultats d'enquêtes, etc.) bien que leur classification, leur format puissent évoluer dans le temps.

Archiver les données

60. À l'heure actuelle, les disques durs des ordinateurs contiennent à la fois des données essentielles au suivi des activités, mais également pléthore d'informations redondantes, obsolètes, voire inutile, sans compter les données personnelles des utilisateurs. Un véritable travail de sélection des données doit être opéré, la solution ne se trouvant pas dans un accroissement continu des capacités de stockage.

61. La duplication des données par stockage redondant sur des supports différents de ceux de l'équipement utilisé (poste de travail fixe, mobile, serveur, etc.) est un des principes de base d'une bonne conservation. Il convient de privilégier un stockage centralisé, la règle du 3-2-1 étant généralement recommandée (3 copies sur 2 supports différents, dont 1 lieu déporté).

62. Se préoccuper de l'archivage des données fait donc partie intégrante d'une bonne gestion des données et est renseignée au sein du PGD/DMP. Dans une logique de préservation, l'archivage se conçoit très en amont d'un projet, dès la création de la donnée. Son objectif est de décrire, documenter, contextualiser les données pour pouvoir ensuite assurer leur diffusion et leur préservation à long terme.

63. La notion d'archivage pérenne concerne l'archivage sur le très long terme (plus de 30 ans) et pose forcément la question de la fin de vie de nos données. Les données sont des archives publiques dès lors qu'elles sont créées au sein d'un établissement public et l'archivage institutionnel est réglementé par la loi et notamment le code du patrimoine. De ce fait, il existe des durées officielles établies par la section des archivistes en universités, rectorats, organismes de recherche et mouvements étudiants (AURORE) au sein de l'association des

archivistes français à travers un « référentiel de gestion des archives de la recherche (Association des archivistes français 2016). Les données doivent faire l'objet d'un tri, d'une sélection, idéalement à la suite d'un échange entre chercheur et archiviste en vue d'une conservation, si nécessaire, aux Archives nationales ou départementales.

Déposer les données dans des entrepôts

64. Un entrepôt de données est un dépôt central informatique contenant des données décrites par un ensemble minimum de métadonnées (titre, licence, créateur, etc.) permettant leur identification (identifiant ou moyen d'accès), leur diffusion et leur réutilisation. Il convient toutefois de garder à l'esprit que vous n'êtes pas maître de la durée de préservation des données dans un entrepôt, cette information de durée dépend de la politique de l'entrepôt. Un entrepôt ne peut donc pas être considéré comme une véritable archive.

65. Il existe beaucoup d'entrepôts de données, certains sont des entrepôts institutionnels (DataSuds, Didomena, etc.), d'autres thématiques (Pangaea pour les données environnementales, Nakala pour les SHS, etc.) ou généralistes (Dryad, Zenodo, etc.) Pour vous aider à trouver et à choisir votre entrepôt, des catalogues sont disponibles[7].

66. Des entrepôts spécifiques peuvent être suggérés (ou imposés) par la revue dans laquelle le chercheur publie un *data paper*, mais aussi par le financeur, le consortium du projet ou votre institution. Il est habituellement recommandé de déposer les données dans un entrepôt de données de confiance (certification CoreTrustSeal[8]).

67. Une nouvelle plateforme nationale fédérée des données de la recherche, Recherche Data Gouv (Ouvrir la Science – Recherche Data Gouv: plateforme nationale fédérée des données de la recherche) a été mise en place au printemps 2022 sous l'égide du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI). Cette plateforme va permettre de favoriser le partage et l'ouverture des données produites par la recherche en fournissant aux chercheurs un entrepôt pluridisciplinaire pour le dépôt des données qui ne trouveraient pas place au sein d'un entrepôt thématique de confiance.

Conclusion

68. La gestion des données associées à un *data paper* répond à un enjeu scientifique indispensable, mais également économique et sociétal qui s'inscrit dans le mouvement de la science ouverte.

7. https://www.re3data.org/, https://fairsharing.org/databases/

8. https://www.coretrustseal.org/.

69. Améliorer les pratiques de gestion des données de la recherche s'avère nécessaire pour assurer la traçabilité de la recherche et garantir l'intégrité scientifique, mais aussi pour rendre accessible, partager et permettre la réutilisation des connaissances produites. Il s'agit d'une préoccupation de toute organisation performante avec l'objectif de disposer de données « FAIR » soit « *Findable, Accessible, Interoperable and Reusable* ». Une telle démarche facilite aussi la production de *data papers* qui à leur tour, contribuent à rendre les données plus FAIR.

70. La FAIRisation des données au sein des structures de recherche est un processus souvent complexe qui nécessite des moyens techniques et humains. Elle doit s'accompagner, d'une part, d'une acculturation de l'ensemble des acteurs de la recherche aux principes de la science ouverte et d'autre part, d'une mise en place d'une organisation destinée à garantir toutes les étapes du cycle de vie des données. Un véritable processus de gouvernance des données est donc nécessaire et doit s'exercer au quotidien par tous les acteurs de la recherche. Ces nouvelles organisations trouvent naturellement leur place dans les démarches d'amélioration continue ou démarches qualité de plus en plus mises en place au sein des structures de recherche[9].

71. Cette gouvernance et ces nouvelles pratiques vont nécessairement impacter les activités des personnels des unités de recherche. Il est essentiel que chaque structure de recherche (unité, service, plateforme, etc.) définisse, formalise, mais également communique la méthodologie adoptée. Les règles ainsi définies devront être largement disponibles et accessibles à tout le personnel de l'unité (sur un intranet, par exemple) avec une diffusion ciblée en direction des nouveaux entrants et du personnel non statutaire.

72. Par ailleurs, il conviendra de se rapprocher des différentes initiatives institutionnelles développées au sein de nos établissements pour accompagner la politique des données de la recherche. Ainsi, la nouvelle plateforme nationale Recherche Data Gouv propose des « ateliers de la donnée » sur l'ensemble du territoire, destinés à conseiller et accompagner les chercheurs tout au long du cycle de vie des données.

9. https://qualite-en-recherche.cnrs.fr/

Le rôle des licences dans la FAIRisation des données

Thomas Margoni, Luca Schirru et Brad Spitz

Introduction

1. Les données sont présentées aujourd'hui comme le nouvel or noir. Cependant, malgré le côté fascinant de l'aspect « valeur » de l'analogie, celle-ci ne s'étend pas au bien lui-même, à sa nature. Le pétrole, comme tout bien déterminé et tangible, est destiné à être consommé, ce qui signifie que son utilisation réduira directement la quantité disponible pour les autres (Ostrom, 1999). Une autre caractéristique des biens tangibles est la facilité avec laquelle on peut exclure. Autrement dit, il est beaucoup plus facile (ou moins onéreux) d'empêcher une personne qui ne paie pas, de profiter d'un bien déterminé (la théorie du « passager clandestin ») lorsque ce bien est matériel (par exemple, une cassette audio), que lorsque l'information est purement immatérielle (par exemple, un fichier MP3). Il est plus aisé d'appréhender les données comme étant un bien public ou, comme l'a catégorisé Benjamin Coriat, comme un savoir commun présentant les caractéristiques suivantes : les données ne sont pas un bien de consommation et ne permettent pas d'exclure (Coriat, 2011 ; Drahos, 2016 ; Fisher, 1987). En raison de ces caractéristiques, il a été soutenu que la gouvernance des données devrait être « orientée non pas vers la *conservation* des ressources », car les données ne s'épuisent pas, mais vers leur *enrichissement* et leur *développement* (Coriat, 2011, p. 13-19).

2. Il est également possible de soutenir que l'on n'a jamais observé un phénomène aussi impactant sur la société que celui que les données ont sur l'économie, les progrès culturels et technologiques. La généralisation des données dans la société (et dans l'économie, la culture, la technologie, etc.) est un exemple récent de la façon dont l'information numérique, exprimée maintenant dans les données, influence les catégories classiques de la connaissance et du pouvoir (Mejias et Couldry, 2019). Si « tout » se transforme en données, réguler les données signifie par extension que l'on doit « tout » réguler, ou du moins beaucoup plus que ce que l'on aurait pu envisager initialement. Prenant apparemment en compte cette fonction changeante dans la régulation des données, les développements législatifs récents de l'UE, tels que le règlement sur la gouvernance des données (Data Governance Act) et le projet de Data Act, semblent proposer un nouveau paradigme pour la régulation des données, un paradigme qui n'est pas (ou pas exclusivement) basé sur des droits de propriété tels que le droit d'auteur, mais sur une sorte de modèle de gouvernance public-privé (Ducuing *et al.*, 2022 ; Baloup *et al.*, 2021).

3. Dans ce chapitre, nous nous référerons aux « données » pour traiter des données à caractère non personnel, qui peuvent, ou non, être protégées par le droit d'auteur et

des droits apparentés. En particulier, nous ne traiterons pas des données personnelles, dans la mesure où celles-ci suivent une approche réglementaire assez différente (par exemple le RGPD), qui n'est pas fondée sur les droits de propriété, mais sur un cadre différent (par exemple, l'autodétermination de l'information). Il est important de relever que les « *data papers* » ne sont pas en eux-mêmes une catégorie discrète dans les sciences juridiques et, d'un point de vue juridique, ils devraient recevoir le même traitement que celui accordé aux autres œuvres littéraires et scientifiques. La seule référence aux « *data papers* » figurant dans la législation analysée se trouve au considérant 28 de la Directive sur les données ouvertes[1], qui indique que « Les États membres peuvent étendre l'application de la présente directive aux données de la recherche rendues accessibles au public [...] sous la forme d'un fichier joint à un article, à un *data paper* ou à un article dans un journal de données (*data journal*) », ce qui ne donne que peu d'éclairage pour interpréter le texte. Cette catégorie particulière de travaux scientifiques appelle toutefois une attention particulière en raison du stockage et de l'utilisation de bases de données, et des quantités importantes de données. En conséquence, ce chapitre analysera comment fonctionne la législation de l'UE sur le droit d'auteur, en insistant sur les données et les bases de données.

4. En ce qui concerne les licences, nous nous concentrerons sur quelques-uns des exemples les plus connus et les plus populaires dans le domaine : *Creative Commons Public Licence* version 4.0 (CCPL) et *Creative Commons Zero* (CC0). Il existe bien sûr de nombreuses autres licences « ouvertes » ; néanmoins, une analyse juridique rigoureuse doit se concentrer sur une étude spécifique de documents juridiques bien identifiés. Compte tenu de la taille et de la portée de ce chapitre, nous utiliserons les licences CCPL et CC0 pour la présente analyse[2]. Enfin, bien que FAIR signifie *Findable, Accessible, Interoperable and Reusable*, compte tenu de l'accent mis dans ce chapitre sur le rôle des licences dans la FAIRisation des données, nous nous concentrerons plus particulièrement sur les principes de *Reusability* et de *Interoperability* (Landi *et al.*, 2020).

5. Dans la première partie de ce chapitre, nous traiterons des principales caractéristiques du FAIR et de ce que cela implique pour les données d'être traçables, accessibles, interopérables et réutilisables. Nous aborderons ensuite brièvement la relation entre les données et le droit d'auteur dans la partie II. Dans la partie III, nous présenterons certains des modèles de licence utilisés pour les données non personnelles, en mettant l'accent sur les licences *Creative Commons* et d'autres régimes permettant l'accès aux données.

1. Directive 2019/1024 du parlement européen et du conseil du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public : https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32019L1024.

2. En ce qui concerne cet aspect, il convient également de noter que la convergence vers un nombre limité de licences standard présente l'avantage d'éviter ou de limiter au problème de « compatibilité des licences » causé par la prolifération des licences. Voir de manière générale, https://en.wikipedia.org/wiki/License_proliferation.

Les données FAIR

6. Les données FAIR (ou FAIR data) sont des données mises à disposition selon quatre grands principes : traçabilité, accessibilité, interopérabilité et réutilisabilité (Wilkinson, Dumontier, Aalbersberg *et al.*, 2016). Comme mentionné ci-dessus, ce chapitre se concentrera sur les données non personnelles et examinera spécifiquement la relation entre les données non personnelles, les licences de droit d'auteur et les principes FAIR. L'analyse appelée « Transformer FAIR en réalité » (Collins et coll., 2018) et d'autres théories importantes, comme l'initiative GO FAIR, sont très utiles pour mieux comprendre comment les données peuvent être conformes aux principes de *findability, accessibility, interoperability* et *reusability*, comme le montre le tableau 1.

7. Parmi les quatre principes mentionnés dans le tableau 1, deux sont au cœur de l'analyse développée pour ce chapitre : la réutilisation et l'interopérabilité. En ce qui concerne le premier, la réutilisation renvoie généralement à la possibilité de réutiliser l'objet sous licence, ce qui, dans le domaine du droit d'auteur, actionne généralement les droits de reproduction, de distribution, d'adaptation et de communication au public. En ce qui concerne l'interopérabilité, il convient de préciser que l'« *Interoperability* » peut prendre différentes formes et renvoyer à divers concepts selon le contexte de l'utilisation (technique, juridique, etc.) (Graber-Soudry *et al.*, 2021). Nous nous concentrerons sur l'interopérabilité juridique (Collins *et al.*, 2018). Nous utilisons notamment la définition de « l'Interopérabilité juridique » adoptée dans l'étude « Legal Interoperability and the FAIR Data Principles », qui est la suivante :

8. > [L'interopérabilité] porte sur la capacité de combiner des ensembles de données provenant de sources multiples sans conflit entre les restrictions liées à chaque ensemble de données (étant précisé qu'une restriction peut intrinsèquement en annuler une autre). [...] L'interopérabilité juridique concerne également les situations dans lesquelles des mesures réglementaires ou politiques restreignent la divulgation de données, ou dans lesquelles des ensembles de données ne peuvent être mis à disposition que dans certains territoires ou sous certaines conditions. Il s'agit par exemple de restrictions juridiques fondées sur le droit de la propriété intellectuelle, la sécurité nationale, la protection d'espèces menacées ou encore une réglementation sur la vie privée telle que le RGPD (Graber-Soudry *et al.*, 2021).

9. Il semble évident que la réutilisation est souvent, peut-être même toujours, une condition préalable à l'interopérabilité. Toutefois, compte tenu des chevauchements potentiels entre l'interopérabilité juridique et la réutilisabilité, sauf indication contraire, cette étude les abordera ensemble, à travers le prisme des droits exclusifs qui peuvent couvrir les données non personnelles lorsque ces données sont protégées par le droit d'auteur ou des droits connexes (Wilkinson, Dumontier, Aalbersberg et al., 2016). Par conséquent, la partie II présentera brièvement ces droits exclusifs, principalement le droit d'auteur et le droit sui generis sur les bases de données.

Bref aperçu des droits et limitations

Droit d'auteur

10. Dans le cadre de la propriété intellectuelle, de nombreux droits exclusifs portent sur divers produits, procédés ou activités. Les signes permettant de distinguer des produits et des services de différentes entreprises peuvent relever du droit des marques (ADPIC, article 15). L'apparence des dessins ou modèles industriels ou des arts appliqués (comme le dessin ou modèle d'un téléphone) peut être protégée par le droit sur les dessins et modèles (ADPIC, article 25). L'invention, qui porte sur un produit ou un procédé, peut être brevetée à condition qu'elle soit « *nouvelle, implique une activité inventive et qu'elle soit susceptible d'application industrielle* », (ADPIC, article 27). Les informations secrètes ayant une valeur commerciale du fait de leur secret peuvent être qualifiées de secrets d'affaires (Directive de 2016 sur les secrets d'affaires). Dans ce chapitre, nous nous concentrerons sur le droit d'auteur qui protège les expressions originales dans les domaines littéraire et artistique. La protection porte à la fois sur les créations classiques, telles que les livres, les poèmes, les brochures, les compositions musicales, les peintures, les sculptures ou les œuvres photographiques, et les productions plus modernes, telles que les logiciels informatiques ou les bases de données.

11. Il est important à ce stade de relever que le terme « données » au sens scientifique ne se traduit pas nécessairement par « données » au sens juridique du terme. En droit d'auteur, les « données » au sens scientifique peuvent être considérées comme de simples faits et données, ou au contraire comme des œuvres de l'esprit, ou encore comme d'autres objets protégés. Il est également important de souligner que les données au sens juridique (ou du moins les *données en tant que telles*) sont souvent, voire systématiquement, exclues du champ de protection par le droit d'auteur. Juridiquement, seules les œuvres de l'esprit (par opposition aux « données ») sont protégées par le droit d'auteur. Il s'agit notamment des créations littéraires et artistiques originales susmentionnées, qui, dans l'UE, doivent être des créations intellectuelles portant l'empreinte de la personnalité de l'auteur, généralement obtenues par les choix libres et créatifs de l'auteur (Margoni, 2016). Il existe une catégorie liée au droit d'auteur, appelée droits connexes ou droits voisins, destinée à accorder une protection – généralement à un degré moindre – aux investissements méritants, qui sont directement ou indirectement liés au processus créatif, mais distincts de celui-ci. C'est notamment le cas des phonogrammes (et, dans l'UE, des films), de la radiodiffusion, des interprétations et exécutions (ces dernières étant cependant souvent considérées comme plus proches du droit d'auteur). Mais, y compris dans cette catégorie, les données en tant que telles ne se voient accorder que peu de protection, voire aucune. D'un autre côté, les « données » au sens scientifique du terme peuvent très bien renvoyer à des éléments protégés par le droit d'auteur ou des droits connexes.

12. Les données en tant que telles sont une catégorie qui pose des difficultés en droit d'auteur. Les données sont généralement considérées comme les éléments constitutifs de la créativité humaine, de la même manière que les lettres de l'alphabet sont les éléments constitutifs d'un poème. Par nature, elles ne sont pas en rivalité avec les biens et peuvent potentiellement créer de la richesse, ce qui pose la question de savoir comment trouver un juste équilibre entre les divers intérêts fondamentaux qui sont liés à l'accès, l'utilisation et la protection des données (Coriat, 2011). En conséquence, les idées, les faits et les données sont souvent présentés par la doctrine comme n'étant pas, en tant que telles, protégées par le droit d'auteur, et comme étant uniquement protégées dans les cas précis et limités définis par les droits connexes (Convention de Berne, article 2). En particulier, la dichotomie idée-expression (et la dichotomie fait-expression qui est liée à la première) constitue l'une des doctrines clés élaborées par les législateurs et la jurisprudence pour trouver un équilibre entre l'accès public et l'initiative privée, la réutilisation créative et le retour sur investissement, la capacité d'apprendre librement à partir des éléments de base du savoir et la nécessité de protéger les activités créatives contre les imitations abusives.

13. *Les bases de données* méritent un examen plus approfondi. Les accords ADPIC prévoient à l'article 10(2) que les bases de données peuvent être qualifiées d'œuvres protégées par le droit d'auteur si la sélection ou la disposition de leur contenu est considérée comme une « *création intellectuelle* ». La seconde partie de l'article 10(2) prévoit que « [c]*ette protection* [...] *ne s'étendra pas aux données ou éléments eux-mêmes* » et « *sera sans préjudice de tout droit d'auteur subsistant pour les données ou éléments eux-mêmes* »[3]. Les bases de données visées à l'article 10(2) de l'Accord sur les ADPIC et, en Europe, dans la première partie de la Directive 96/9/CE (« Directive sur les bases de données »), sont communément appelées « bases de données originales » dans la mesure où, pour être protégée par le droit d'auteur, leur disposition ou leur sélection doit être originale. À titre d'exemple, le fait d'organiser une base de données uniquement dans l'ordre chronologique ou alphabétique n'est pas suffisant pour que la base soit éligible à la protection par le droit d'auteur, car la structuration des données selon des principes purement chronologiques ou alphabétiques ne relève pas de la création intellectuelle. C'est le cas même lorsque la création de la base de données a nécessité un investissement substantiel en temps, en argent ou en efforts. Toutefois, une base de données dont la sélection ou l'arrangement suit les choix originaux de l'auteur (la sélection ou l'arrangement n'étant donc pas alphabétique ou chronologique, mais basé sur certains choix originaux faits par l'auteur) peut très bien être protégée par le droit d'auteur, même si cela n'a pas nécessité d'efforts particuliers. Le droit d'auteur protège l'originalité et non les efforts ou les investissements. Il convient toutefois de souligner que, dans ce cas de figure, c'est la structure de

3. En d'autres termes, s'il s'agit d'une base de données de photographies, les droits d'auteur sur les photographies subsistent – et sont gérés, par exemple, par le photographe – quel que soit le traitement juridique spécifique accordé à la base de données.

la base de données qui est originale et donc protégée, et non le contenu.

14. Ainsi, par exemple, en l'absence de tout autre droit, il serait licite de réutiliser les éléments (les données) de la base de données sans enfreindre le droit d'auteur, à condition toutefois que la structure (la sélection ou l'arrangement) ne soit pas reproduite. Les annuaires téléphoniques en sont l'exemple parfait. Les « pages blanches » qui répertorient tous les utilisateurs de téléphones par ordre alphabétique peuvent tout à fait être des bases de données, sans pour autant être protégées par le droit d'auteur, puisque la sélection (tous les utilisateurs de téléphone) et la disposition (alphabétique) ne sont pas une création intellectuelle propre à l'auteur. Au contraire, les « pages jaunes », qui utilisent souvent une sélection limitée (seulement certains utilisateurs et/ou zones) et une disposition également limitée (par catégorie, par taille, etc.) peuvent atteindre le niveau d'originalité requis. Cela étant, même dans ce dernier cas, ce sont les critères de sélection et/ou d'arrangement (la structure) qui sont protégés par le droit d'auteur, et non les noms et numéros de téléphone (le contenu). Nous verrons dans la section suivante qu'au sein de l'UE un droit *sui generis* sur les bases de données offre une couche de protection supplémentaire dans ce cas de figure.

15. Une autre caractéristique fondamentale du droit d'auteur est la territorialité, dans la mesure où, malgré l'existence d'accords internationaux et régionaux visant à coordonner et à harmoniser les règles, le droit d'auteur reste un droit régi par la législation nationale. À titre d'illustration de ce principe, alors que les règles internationales fixent la durée de protection du droit d'auteur à un minimum de 50 ans *post-mortem*, le droit national (ou de l'UE) a la possibilité d'étendre la protection au-delà, en l'occurrence à 70 ans *post-mortem*, comme c'est le cas aux États-Unis et dans l'UE. Une fois la durée de protection écoulée, les œuvres tombent dans le domaine public et sont libres d'être utilisées par tous (bien que les droits moraux puissent perdurer).

16. Le droit d'auteur porte sur les droits spécifiques d'exploitation économique (généralement les droits de reproduction, d'adaptation, de distribution et de communication au public), ainsi que certaines limitations ou exceptions. Les faits divers et les nouvelles du jour sont quelques-uns des exemples de ce qui est expressément exclu du champ d'application du droit d'auteur (Convention de Berne, article 2.8 ; Margoni et Kretschmer, 2022). Les limitations et exceptions, de même que la durée de la protection, peuvent varier d'une législation nationale à une autre, mais certaines d'entre elles s'imposent aux signataires de certains accords internationaux, par exemple l'exception de citation dans le cadre de la Convention de Berne (article 10.1). Les directives de l'UE peuvent également prévoir des exceptions obligatoires ou facultatives à transposer dans le droit national, comme c'est le cas de la Directive 2001/29/CE qui prévoit une exception obligatoire et 20 exceptions et limitations facultatives, ou encore de la Directive (UE) 2019/790 sur le droit d'auteur et les droits voisins

dans le marché unique numérique, qui prévoit des exceptions (obligatoires) pour la fouille de textes et de données (TDM) (Margoni et Kretschmer, 2022).

Droit *Sui Generis* sur les Bases de Données

17. Au sein de l'UE, la Directive sur les bases de données prévoit en outre un droit *sui generis* pour les bases de données, indépendant et cumulable avec le droit d'auteur (Directive sur les bases de données, article 7 [1]). Alors que le droit d'auteur est accordé au titre de la disposition ou de la sélection originale du contenu, « l'objet de ce droit *sui generis* est d'assurer la protection d'un investissement dans l'obtention, la vérification ou la présentation du contenu d'une base de données pour la durée limitée du droit » (Directive sur les bases de données, considérant 40). Le considérant 40 de la Directive sur les bases de données précise que l'investissement « peut consister dans la mise en œuvre de moyens financiers et/ou d'emploi du temps, d'efforts et d'énergie ». Les considérants de la Directive sur les bases de données (par exemple, le considérant 19) et les décisions de la CJUE dans les affaires *Fixtures Marketing*[4] et *British Horseracing Board*[5] donnent des précisions pour déterminer ce qui peut constituer un investissement portant sur l'obtention, la vérification et la présentation du contenu de la base de données.

18. Une précision importante dans la décision *British Horseracing Board* fait référence au véritable objectif de la Directive sur les bases de données, qui est (était ?) de « stimuler la mise en place de systèmes de stockage et de traitement d'informations existantes, et non la création d'éléments susceptibles d'être ultérieurement rassemblés dans une base de données »[6]. En d'autres termes, la Directive sur les bases de données ne vise pas à favoriser la création – et la protection juridique – des données elles-mêmes, mais uniquement des systèmes de stockage et de traitement des données.

19. Un point important, et souvent mal compris, concernant les bases de données, est que les données contenues dans une base de données ne sont pas protégées par le droit d'auteur (éventuel) sur la base de données. Les données ne sont pas non plus directement protégées par le droit *sui generis*. En réalité, une donnée isolée ou des quantités non substantielles de données n'entrent pas dans le champ d'application du droit *sui generis*. Ce n'est que lorsqu'un *investissement substantiel* dans l'obtention, la vérification ou la présentation des données a été réalisé que des *quantités substantiellement importantes* de données dans la base de données sont protégées contre leur extraction et réutilisation. Les données créées (ou les données de meilleure qualité lorsqu'un investissement substantiel a été réalisé dans la phase de création) ne sont pas non plus protégées, et ce, afin d'éviter de créer des situations anticoncurrentielles sur les marchés de l'information

4. Affaire C-338/02 *Fixtures Marketing Ltd v Svenska Spel AB* [2004] ECR I-10497, paras 27-28.
5. Affaire C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* [2004] ECR I-10415, para 31.
6. Id.

(telles que les bases de données dites de source unique). Il ne peut y avoir de protection que dans la mesure où les données ont été collectées (et qui sans doute existaient déjà) et que l'investissement est lié à la constitution de ces données (Hugenholtz, 2018).

20. Bien entendu, les différents éléments constituant la base de données peuvent en même temps être protégés par le droit d'auteur. Ce sera par exemple le cas d'une base de données d'articles de journaux, lesquels bénéficient généralement d'une protection par le droit d'auteur. Cela peut créer une situation potentiellement complexe dès lors que, pour une même base de données, il peut y avoir trois niveaux de protection différents : le droit d'auteur sur la structure de la base de données, le droit *sui generis* sur l'investissement substantiel, et le droit d'auteur sur les éléments individuels (les articles de journaux) de la base de données. On ne peut exclure l'existence de situations dans lesquelles chacune de ces trois couches de protection suit des voies différentes s'agissant de la titularité, de la protection et de la réutilisation, dans la mesure où l'auteur de la base de données, le producteur titulaire du droit sui generis et les auteurs des articles de journaux peuvent très bien être des personnes différentes, chacune titulaire d'un droit lié à sa situation.

21. Par conséquent, il devient primordial que les licences, et en particulier celles qui accordent des droits de réutilisation larges destinés à faire progresser les principes FAIR, prennent bien en compte cette complexité juridique.

Niveaux réglementaires supplémentaires

22. Il convient de souligner que les entreprises à forte intensité de données et les organismes de recherche doivent faire face non seulement aux droits existants protégeant les bases de données (droit d'auteur et droit *sui generis*), mais également à plusieurs autres niveaux de protection, qu'ils soient d'origine législative, contractuelle ou technologique (Souza, Schirru et Alvarenga, 2020 ; Souza, 2020). La couche technique est particulièrement bien illustrée par les mesures techniques de protection (MTP) et la gestion des droits numériques (DRM) (Souza, Schirru et Alvarenga, 2020 ; Souza, 2020). La couche contractuelle est illustrée non seulement par les licences, qui seront discutées infra, mais aussi par les « Conditions d'Utilisation » qui, au-delà des droits sous-jacents spécifiques, peuvent être utilisées pour limiter ou interdire l'accès et/ou l'utilisation des données (Souza, Schirru et Alvarenga, 2020 ; Souza, 2020).

23. Ce que nous avons présenté semble déjà constituer un cadre complexe de règles fondées sur l'exclusivité, avec des effets incertains sur la recherche et l'innovation (Souza, Schirru et Alvarenga, 2020 ; Souza, 2020 ; Dosi et Stiglitz, 2013) ; pourtant, l'idée d'étendre ou de créer *ex novo* un droit protégeant les données générées informatiquement (qui semblent actuellement largement exclues en application de la dichotomie création *versus* obtention) était, jusqu'à récemment, en discussion au niveau législatif au sein de l'UE (Hugenholtz, 2018). Bien que ces données puissent être utiles pour certaines industries à forte intensité de

données, elles peuvent échapper à la protection par la propriété intellectuelle. Pas uniquement parce qu'elles sont générées par des machines, mais également parce qu'elles seront probablement de simples faits et données qui, comme nous l'avons vu *supra*, ne relèvent pas, à juste titre, du droit d'auteur (Convention de Berne, article 2.8 ; ADPIC, article 9[2]). Même si l'on accepte de mettre pour un instant de côté la logique de la propriété intellectuelle, dans une perspective d'organisation industrielle, il est loin d'être certain que la restriction à l'accès à ces données par le biais de droits de propriété constitue la meilleure solution, et ce, en raison de l'importance des effets anticoncurrentiels des monopoles de données sur le commerce et l'innovation (Drexl *et al.*, 2019).

24. En outre, il semble avoir été bien démontré qu'il n'est pas nécessaire de créer un nouveau droit de propriété pour encourager la production de données, non seulement en raison de l'absence de doctrine démontrant une défaillance manifeste du marché (Drexl *et al.*, 2017), mais encore parce que cela irait à l'encontre de la logique fondamentale de l'économie des données (Gärtner et Brimsted, 2017 ; Kerber, 2016). Sur ce dernier point, Drexl *et al.* (2017, p. 6) expliquent qu'« *u*n nouveau système de droits de propriété ne devrait être créé que si un tel droit améliore le fonctionnement de l'économie des données »[7]. Les évolutions législatives les plus récentes de l'UE dans ce domaine (dont certaines sont encore à l'état de projet) prennent une autre direction, comme le montre l'article 35 du projet de règlement sur les données qui limite la portée du droit *sui generis* en précisant qu'il ne s'applique pas aux données IoT (*Internet of Things*). On pourrait soutenir qu'il ressort des développements législatifs et politiques les plus récents de l'UE, que la direction qui semble être prise est moins celle de la propriété des données, que celle d'espaces de données européens communs (Ducuing *et al.*, 2022).

Licences ouvertes

25. Lorsque les « données » bénéficient de la protection du droit d'auteur et/ou du droit *sui generis*, comme expliqué ci-dessus, il existe deux principales possibilités pour les réutiliser : (i) la réutilisation dans le cadre des autorisations légales, par exemple, les exceptions et limitations au droit d'auteur prévues par la loi ; ou (ii) la réutilisation dans le cadre d'une autorisation contractuelle, par exemple les licences de droit d'auteur.

26. En ce qui concerne le point (ii), le problème est que les licences doivent souvent être négociées individuellement entre les utilisateurs (bénéficiaires de la licence) et les titulaires de droits (donneurs de licence), situation

7. Voir également Kerber (n 54) 3 (« *En conclusion, cet article établira – en se fondant sur nos connaissances préliminaires actuelles – qu'un nouveau droit de propriété intellectuelle sur les données n'est pas nécessaire (notamment en raison de l'absence de difficultés liées à l'incitation à produire et à analyser des données). Au contraire, la création d'un droit peut même être dangereuse pour l'innovation et la concurrence dans l'économie numérique, en conduisant à des difficultés d'interprétations juridiques sérieuses, à la monopolisation de l'information et à des entraves à la libre circulation des données qui sont essentielles pour l'économie numérique*. »)

qui est souvent affectée par des déséquilibres de pouvoir dans la négociation du contrat (il suffit de penser aux contrats BtoC et au droit de la consommation destiné à protéger la partie faible). Un autre problème courant est la difficulté de négocier des conditions contractuelles exécutoires, multi-territoriales et équitables, en particulier lorsqu'il s'agit principalement d'obtenir des autorisations permettant la réutilisation et l'interopérabilité, comme l'exigent les principes FAIR.

27. Les licences dites « ouvertes » visent à aider les créateurs, les utilisateurs, les développeurs et les producteurs (« prosommateurs ») à créer un marché plus équitable, transparent et efficace en offrant des licences publiques « pré-packagés » et souvent modulaires, destinées à tous. Voyons comment cela se présente dans le projet Creative Commons.

Creative Commons

28. *Wikipedia*, *Flickr*, *Youtube*, *Google Images*, et autres... Tous fonctionnent, d'une manière ou d'une autre, avec ou grâce aux licences *Creative Commons*[8]. Les moteurs de recherche d'images de *Google Images* et *Flickr* permettent aux utilisateurs de filtrer leurs résultats pour afficher le contenu pouvant, par exemple, être utilisé à la fois à des fins commerciales et non commerciales, à la seule condition de citer correctement. Il en va de même pour la recherche de vidéos sur *YouTube*. Les publications de *Wikipédia* sont sous licence CC BY-SA 3.0.

29. Au mépris du paradigme « tous droits réservés », et au profit de « certains droits réservés »[9], *Creative Commons* a principalement pour but « d'augmenter la quantité de créations sous licence ouverte en "commun" – l'ensemble des œuvres librement disponibles pour un usage licite, le partage, la réutilisation et le remixage » (CC, 2022a)[10].

30. Il est important de relever que les licences *Creative Commons* ont des mécanismes intelligents intégrés dans la licence pour permettre une adaptation à la loi nationale sur le droit d'auteur (n'oubliez pas la territorialité... Voir ci-dessus). Ainsi, dans le cas de la renonciation prévue dans la licence CC0, le code juridique prévoit explicitement que la renonciation au droit d'auteur est limitée à « ce qu'autorise la loi applicable, sans violation de celle-ci (...) » (CC, 2022e). Dans les juridictions où les renonciations au droit d'auteur ne sont pas envisagées, CC0 fonctionnera comme une licence avec un effet juridique aussi proche que possible de celui recherché. De même, dans les pays où les droits moraux existent (pratiquement tous les pays du monde à l'exception

8. Sur la relation entre l'auteur, les utilisateurs et le rôle des licences CC, voir S Dusollier, « The Master's Tools v. The Master's House : Creative Commons v. Copyright » (2006) 29 Columbia Journal of Law & Arts, 271 : https://papers.ssrn.com/sol3/papers.cfm?abstract_id =2186187, consulté le 12/02/2025.

9. Voir MW Carroll, Creative Commons as Conversational Copyright (2007) in PK Yu (ed) *Intellectual Property and Information Wealth: Issues and Practices in the Digital Age* (Praeger, 2007), Villanova Law/Public Policy Research Paper No. 2007-8 : https://ssrn.com/abstract=978813, consulté le 21 déc. 2022.

10. Sur le rôle des licences CC dans la promotion du libre accès, voir T Margoni, D Peters, « Creative Commons Licenses : Empowering Open Access » (2016) : https://ssrn.com/abstract =2746044, consulté le 21 déc. 2022.

partielle des États-Unis) et ne peuvent faire l'objet d'une renonciation (les droits moraux sont incessibles, mais on peut y renoncer dans certains droits nationaux), la renonciation CC0 sera, par exemple, limitée aux droits patrimoniaux[11]. Les licences *Creative Common* ont évolué au fil du temps et la version actuelle 4.0 inclut non seulement le droit d'auteur en tant que tel, mais aussi les droits voisins et, ce qui est important pour la présente analyse, le droit *sui generis*. Cela signifie que, lorsqu'une base de données (telle que décrite ci-dessus) est concédée sous licence CCPL BY, les conditions de la licence s'appliquent à l'ensemble du matériel « sous licence », sauf indication contraire. Cela signifie qu'il est fondamental (et il s'agit d'une exigence de la licence) que les donneurs de licence soient juridiquement en mesure d'appliquer la licence aux trois couches (c'est-à-dire le droit d'auteur sur la structure de la base de données, le droit *sui generis* sur des quantités substantielles de données, et le droit d'auteur sur les éléments de la base de données), ou à défaut (c'est-à-dire en cas d'« indication contraire ») d'identifier quelles parties du matériel sous licence sont exclues du CCPL[12].

31. De plus, l'Attribution (BY), qui était initialement une option (version 1.0), est depuis devenue un élément obligatoire, y compris pour les concédants de licence basés aux États-Unis[13]. Enfin, les licences CC sont destinées à toutes les créations de droits d'auteur à l'exception des logiciels, pour lesquels les licences FLOSS sont utilisées (et ont d'ailleurs inspiré le CC) depuis de nombreuses années.

Principes FAIR et licences CC

32. Les principes Trouvable (*Findability*) et Accessible (*Accessibility*) sont rendus effectifs grâce aux icônes et outils de recherche qui permettent de trouver et d'identifier facilement les conditions des licences (moteurs de recherche, plug-ins des navigateurs, etc.) et sont capables de trouver et de lire les métadonnées de licence des œuvres sous licence CC. Cela est rendu possible par la mise à disposition des licences CC en trois langues. Un contrat lisible par l'homme, avec un résumé des principales obligations de l'utilisateur en application de la licence ; un langage juridique compréhensible pour un juriste, avec les conditions de la licence ; et des métadonnées lisibles par machine (CC, 2022b). Les métadonnées sont celles de *Creative Commons Rights Expression Language* (CC REL), qui « peuvent être intégrées dans différents types de fichiers » (CC Wiki, 2021). En d'autres termes, le fait d'appliquer correctement les métadonnées de licence

11. Voir, par exemple, Brésil, Loi n° 9.610 du 19 février 1998 (Loi sur le droit d'auteur et les droits voisins), art. 27 (« *Les droits moraux sont inaliénables et irrévocables*. »).

12. « *Précisez correctement ce que vous donnez en licence* » (« *Pour chaque œuvre, il existe différents éléments ; par exemples des textes, images, de la musique. Pensez à marquer ou à indiquer clairement dans un avis ceux qui sont couverts par la licence*. »), https://wiki.creativecommons.org/wiki/Considerations_for_licensors_and_licensees#Specify_precisely_what_it_is_you_are_licensing.

13. Voir, par exemple, « CC Attribution 4.0. International', Sec 2(b)(1)(2) (*Creative Commons)* (« *Autres droits. 1. Les droits moraux, tel que le droit à l'intégrité de l'œuvre, ne sont pas accordés par la présente Licence publique, ni le droit à l'image, ni le droit au respect de la vie privée, ni aucun autre droit de la personnalité ou apparenté ; cependant, dans la mesure du possible, le Donneur de licence renonce et/ou accepte de ne pas faire valoir les droits qu'il détient de manière à Vous permettre d'exercer les Droits accordés par la licence. 2. Le droit des brevets et le droit des marques ne sont pas concernés par la présente Licence publique*. ») : https://creativecommons.org/licenses/by/4.0/legalcode, consulté le 21 novembre 2022.

(qui incluent souvent des métadonnées pertinentes sur l'œuvre) au fichier constitue un moyen puissant, mais pas toujours pleinement exploité, pour assurer à la fois Trouvable et Accessible. Le moteur de recherche https://search.creativecommons.org est un exemple de mise en œuvre pratique de ces fonctionnalités.

33. En choisissant d'utiliser un mécanisme *License Chooser* (CC, 2022c), le donneur de licence peut opter pour différentes combinaisons d'utilisations pouvant être faites d'une œuvre sous la licence sélectionnée. La licence la plus permissive est le CC-BY, qui « permet [aux licenciés] de distribuer, remixer, adapter et construire à partir du matériel sur n'importe quel support ou format, tant que le créateur est crédité » (CC, 2019), à la fois pour des utilisations commerciales et non commerciales. Toutes les autres licences peuvent présenter une combinaison de l'exigence d'attribution (BY) avec des exigences supplémentaires, telles que : NC – Utilisation non commerciale[14] ; SA – Partager dans les mêmes conditions[15] ; et ND – Aucune adaptation autorisée[16]. CC offre un outil supplémentaire, à savoir la dérogation CC0, « qui permet aux créateurs de renoncer à leurs droits d'auteur et de mettre leurs œuvres dans le domaine public mondial ». Enfin, l'outil *Public Domain Mark* est simplement un outil permettant d'identifier les œuvres qui sont déjà dans le domaine public (CC, 2022d).

34. En conséquence, tous les outils CC permettent de copier et de redistribuer le contenu. Les licences les plus ouvertes, CCPL BY et CC0, permettent également la création et la distribution d'œuvres dérivées pour toutes les finalités. La plus restrictive, CCPL NC-ND n'autorise l'utilisation du matériel qu'à des fins non commerciales et ne permet pas la création d'œuvres dérivées. En d'autres termes, il est possible d'affirmer que la licence CC0 et la licence CCPL BY de base assurent la Réutilisabilité le plus largement possible. Lorsque des clauses supplémentaires sont appliquées à la licence CCPL BY, telles que NC et/ou ND, la réutilisation est limitée de diverses manières, mais ne disparaît pas totalement. Par conséquent, en regardant le R d'un point de vue juridique, les licences CC sont, d'une manière générale, une bonne indication de la Réutilisabilité, mais il faut vérifier quelle est la licence précisément utilisée.

35. En ce qui concerne l'exigence d'Interopérabilité (*Interoperability*), d'un point de vue juridique, il est important de vérifier si la licence permet de combiner le matériel sous-jacent avec des œuvres sous d'autres licences. Une fois de plus, ce sont les outils les plus ouverts, CC0 et CCPL, qui accordent le niveau d'interopérabilité le plus élevé, car ils permettent la création et la distribution d'œuvres dérivées (qui sont normalement définies comme des œuvres dérivées ou basées sur le matériel original sous licence) sans restriction (pour CC0) à part

14. *À propos des Licences CC* (*Creative Commons*, 2019) (« *Seuls des usages non commerciaux des œuvres sont permis* »), https://creativecommons.org/about/cclicenses/.
15. *À propos des Licences CC* (*Creative Commons*, 2019) (« *Les adaptations doivent être partagées dans les mêmes conditions* »), https://creativecommons.org/about/cclicenses/.
16. *À propos des Licences CC* (*Creative Commons*, 2019) (« *Aucune œuvre dérivée ou adaptation n'est permise* ») https://creativecommons.org/about/cclicenses/.

celles liées à l'obligation de citer la source (CCPL BY). La clause SA de la CCPL (similaire à l'élément copyleft des licences logicielles telles que GPL) exige que les œuvres dérivées soient concédées sous la même licence ou sous une licence compatible. Cela signifie que les œuvres CCPL BY-SA ne peuvent être combinées avec d'autres œuvres que sous une licence CCPL BY-SA ou une licence compatible. Il convient de préciser que l'interopérabilité n'implique pas nécessairement la possibilité de faire une adaptation à partir du matériel. Il existe des situations dans lesquelles deux œuvres peuvent « interopérer », tout en conservant leur singularité et leur autonomie. Par exemple, deux ensembles de données qui communiquent entre eux, mais qui ne sont pas fusionnés, pourraient sans doute être tous deux basés sur une licence CCPL BY-ND, laquelle ne permet pas la création d'œuvres dérivées. Dans ce cas, les deux ensembles de données peuvent « communiquer entre eux », mais *a priori* sans être fusionnés pour former un nouvel ensemble. Cela serait toutefois possible si les ensembles de données étaient chacun sous une licence CC BY ou chacun sous une licence CC BY-SA. Si l'un était sous CC BY et l'autre sous CC BY-SA, il serait nécessaire de publier le résultat sous une licence CC BY-SA, etc[17].

36. Nous relevons en conclusion qu'alors que *Findability* et *Accessibility* ne semblent pas être liés à l'outil CC utilisé (dans le sens où tous les outils CC offrent le même niveau de traçabilité et d'accessibilité), *Reusability* et plus encore *Interoperability* seront affectées par le choix des outils CC. Dans le tableau 1, nous montrons comment les éléments des licences CC reflètent les exigences FAIR.

17. Sur les FLOSS, voir, par exemple, T. Margoni, M. Perry, « Free-Libre Open Source Software as a Public Policy Choice » (2010) 3 (3,4) International Journal on Advances in Internet Technology, 212.

Autres régimes d'accès aux données

37. La question de la réutilisation des données ne peut se limiter à une analyse fondée uniquement sur les droits de propriété. Comme nous l'avons vu, les droits de propriété en général, et le droit d'auteur en particulier, ne sont pas des outils conçus pour réguler correctement les données. En réalité, la protection des données par la propriété a été exclue – ou strictement encadrée – pour des raisons évidentes de politique industrielle et d'innovation. Certaines « données » peuvent trouver une place dans la maison du droit d'auteur, mais uniquement dans le cadre des conditions restrictives présentées ci-avant. Cela étant, la contradiction apparente d'une catégorie (les données), qui acquiert une valeur économique et sociale sans précédent, mais qui ne peut pas être correctement régie par les outils juridiques (le droit d'auteur) dont nous estimons souvent – à tort – qu'ils devraient la régir, peut avoir contribué à orienter le législateur de l'UE vers une voie différente. Nous n'examinerons pas ces régimes d'accès alternatifs en détail, car cela dépasserait la portée – et l'espace – du présent chapitre. Néanmoins, il est utile de les présenter, même brièvement.

Principe	Définition (https://www.go-fair.org/fair-principles/)	Description *(Transformer FAIR en réalité : rapport final et plan d'action de FAIR Data, Commission européenne, 2018, 19-20 (notes de bas de page supprimées)*	Creative Commons Public License (CCPL) 4.0
Traçabilité	« F1. Les (méta)données se voient attribuer un identifiant global unique et persistant F2. Les données sont décrites avec des métadonnées riches (définies par R1 ci-dessous) F3. Les métadonnées incluent clairement et explicitement l'identifiant des données qu'elles décrivent F4. Les (méta)données sont enregistrées ou indexées dans une ressource consultable »	**« Les données sont Traçables lorsqu'elles sont décrites par des métadonnées suffisamment riches et enregistrées ou indexées dans une ressource consultable connue et accessible aux utilisateurs potentiels. De plus, un identificateur unique et persistant devrait être attribué de manière que les données puissent être référencées et citées sans équivoque dans les communications de recherche. »**	CCPL est exprimé en langage juridique, humain et machine (métadonnées). Les œuvres sous licence appropriée sont donc facilement trouvables, bien que CC ne fournisse pas en soi un référentiel par défaut. Cependant, Openverse (https://wordpress.org/openverse) permet de rechercher 600 millions d'œuvres sous licence CC et autres licences ouvertes.
Accessibilité	« A1. Les (méta)données sont récupérables par leur identifiant à l'aide d'un protocole de communication standardisé A1.1 Le protocole est ouvert, gratuit et universellement applicable A1.2 Le protocole prévoit une procédure d'authentification et d'autorisation, le cas échéant A2. Les métadonnées sont accessibles, même lorsque les données ne sont plus disponibles. »	**« Les objets de données accessibles peuvent être obtenus par les humains et les machines avec l'autorisation appropriée et grâce à un protocole bien défini et universellement implémentable. En d'autres termes, toute personne disposant d'un ordinateur et d'une connexion Internet devrait pouvoir accéder au moins aux métadonnées. Il est important de souligner qu'Accessible dans FAIR ne signifie pas Ouvert sans contrainte. L'accessibilité signifie que l'homme ou la machine reçoit – par le biais de métadonnées – les conditions précises dans lesquelles les données sont accessibles et que les mécanismes et protocoles techniques d'accès aux données sont mis en œuvre de manière que les données et/ou les métadonnées puissent être consultées et utilisées à grande échelle, par des machines, sur le Web. »**	Pour des raisons similaires à celles mentionnées ci-dessus, les œuvres CCPL sont très souvent (mais pas nécessairement) accessibles via Internet et les conditions d'accès et de réutilisation sont clairement établies dans les métadonnées, dans l'avis de droit d'auteur lisible par l'homme ainsi que dans la licence complète.
Interopérabilité	« I1. Les (méta)données utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances. I2. Les (méta)données utilisent des vocabulaires qui suivent les principes FAIR I3. Les (méta)données comprennent des références qualifiées à d'autres (méta)données »	**« Les données et métadonnées interopérables sont décrites dans les principes FAIR comme celles qui utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances. Elles utilisent des terminologies qui suivent elles-mêmes les principes FAIR et incluent des références qualifiées à d'autres données ou métadonnées. Ce que cela décrit est l'interopérabilité sémantique. […] Il ne s'agit pas seulement de sémantique, mais aussi d'interopérabilité technique et juridique. L'interopérabilité technique signifie que les données et les informations connexes sont codées à l'aide d'une norme qui peut être lue sur tous les systèmes applicables. Dans FAIR, l'interopérabilité juridique relève du principe selon lequel les données doivent être "réutilisables". »**	L'interopérabilité aux fins de la concession de licences peut être conceptualisée comme la capacité de deux licences différentes à fonctionner ensemble et à permettre aux œuvres sous-jacentes d'être licitement combinées, réutilisées et repartagées. Les licences ouvertes comme CCPL ont normalement des degrés élevés d'interopérabilité ; néanmoins, lorsque deux licences sont fortement copyleft, cela peut conduire à une incompatibilité et donc à un manque d'interopérabilité (l'exemple habituel est la GPLv2 et la GPLv3[18] qui sont toutes deux fortement copyleft et ne permettent pas la redistribution d'œuvres dérivées). CCPL utilise un langage spécifique (ou « licence équivalente ») qui favorise l'interopérabilité.

Tableau 1. Comment les licences CC reflètent les exigences FAIR

18. « GPL-Compatible Free Software Licenses » (GNU Operating System, 12 janvier 2022) : https://www.gnu.org/licenses/license-list.html#GPLCompatibleLicenses, consulté le 21 décembre 2022.

Réutilisabilité	« R1. Les (méta)données sont décrites en détail avec une pluralité d'attributs précis et pertinents R1.1. Les (méta)données sont publiées avec une licence d'utilisation des données claire et accessible R1.2. Les (méta)données sont associées à une provenance détaillée R1.3. Les (méta)données répondent aux normes communautaires pertinentes pour le domaine »	**« Pour que les données soient Réutilisables, les principes FAIR réaffirment la nécessité de métadonnées et d'une documentation riches qui répondent aux normes communautaires pertinentes et fournissent des informations sur la provenance. Cela couvre la façon dont les données ont été créées (par exemple, les protocoles d'enquête, les processus expérimentaux, les informations sur l'étalonnage et l'emplacement des capteurs) et les informations sur les processus de réduction ou de transformation des données pour rendre les données plus utilisables, compréhensibles ou "prêtes pour la science". Comme le montre l'exemple de l'étude de cas DOBES (Fig. 3), les formats ouverts approuvés par la communauté jouent également un rôle clé dans la réutilisabilité. La capacité des humains et des machines à évaluer et à sélectionner les données sur la base de critères relatifs aux informations de provenance est essentielle à la réutilisation des données, en particulier à grande échelle. La réutilisation exige également que les données soient publiées avec une "licence d'utilisation des données claire et accessible": en d'autres termes, les conditions dans lesquelles les données peuvent être utilisées doivent être transparentes pour les humains et les machines. »**	La réutilisation à des fins juridiques semble faire référence à une étape qui précède l'interopérabilité. La réutilisabilité fait référence à la possibilité de réutiliser le matériel sous licence (grâce à une licence qui autorise la réutilisation, c'est-à-dire les reproductions, les adaptations et les redistributions ou les communications publiques). CCPL autorise toujours ces actes, bien que certaines conditions puissent les limiter (NC ou ND). La condition BY garantit que la provenance de l'œuvre, la paternité et la chaîne de modifications sont préservées.

Tableau 1 [suite]. Comment les licences CC reflètent les exigences FAIR

38. *Données de recherche*. La Directive sur les données ouvertes (DDO) définit les données de recherche comme « des documents sous forme numérique, autres que des publications scientifiques, qui sont recueillis ou produits dans le cadre d'activités de recherche scientifique et qui sont utilisés comme preuves dans le processus de recherche ou qui sont communément acceptés dans le milieu de la recherche comme nécessaires pour valider les constatations et les résultats de la recherche » (article 2[9] DDO). La directive impose aux États membres de l'UE de veiller, par le biais de politiques nationales de libre accès, à ce que les données de la recherche financée par des fonds publics soient librement disponibles (via des « politiques de libre accès »), conformément au principe d'« ouverture par défaut », et compatibles avec les principes FAIR (article 10 DDO). L'article 10 précise en outre que les données de recherche doivent pouvoir être réutilisées à des fins commerciales et non commerciales dans la mesure où des chercheurs, des organismes exerçant une

activité de recherche ou des organismes finançant une activité de recherche les ont déjà rendues publiques par l'intermédiaire d'une archive ouverte institutionnelle ou thématique.

39. *Ensembles de données de forte valeur.* Le même DDO définit les ensembles de données de forte valeur comme des documents dont la réutilisation est associée à d'importantes retombées positives au niveau de la société, de l'environnement et de l'économie, en particulier parce qu'ils se prêtent à la création de services possédant une valeur ajoutée, d'applications et de nouveaux emplois décents et de grande qualité, ainsi qu'en raison du nombre de bénéficiaires potentiels des services et applications à valeur ajoutée fondés sur ces ensembles de données (article 2[10] DDO). Les ensembles de données de grande valeur doivent être mis à disposition en vue d'une réutilisation dans un format lisible par machine, au moyen d'une API appropriée et, le cas échéant, sous forme de téléchargement de masse et gratuitement pour l'utilisateur final. Le chapitre V du DDO prévoit un ensemble de règles pour cette importante catégorie de données.

40. *Données IoT.* Le projet de règlement sur les données définit les données de l'*Internet of Things* (IoT) comme toute représentation numérique d'actes, de faits ou d'informations et toute compilation de ces actes, faits ou informations, notamment sous la forme d'enregistrements sonores, visuels ou audiovisuels produits par un objet mobilier corporel, y compris lorsqu'ils sont incorporés dans un bien immeuble, qui obtient, génère ou recueille des données concernant leur utilisation ou leur environnement, qui est en mesure de communiquer des données par l'intermédiaire d'un service de communications électroniques accessible au public et dont la fonction première n'est pas le stockage et le traitement de données. Les utilisateurs de ces produits ont le droit d'accéder gratuitement et en temps réel et de partager ces données avec des tiers sous certaines conditions (proposition de règlement sur les données, articles 3-5).

Conclusion

41. Les données de recherche, les ensembles de données de forte valeur et les données de l'IoT ne sont que quelques-uns des exemples d'une nouvelle approche du droit de l'UE en matière de réglementation de l'accès aux données et de leur réutilisation, fondée sur un paradigme différent de celui de la propriété décrite ci-dessus. Ce nouveau paradigme basé sur la gouvernance, qui a déjà été étudié (Ducuing *et al.*, 2022), suggère un nouveau modèle de gouvernance public-privé des données fondé sur le projet ambitieux d'espaces européens communs de données[19]. Néanmoins, les règles plus traditionnelles fondées sur le droit d'auteur ou les droits connexes portant sur la circulation des connaissances resteront, du moins pour le moment, le paradigme dominant – mais qui n'est désormais plus le seul.

19. 'A European Strategy for data' (European Commission, 2022) : https://digital-strategy.ec.europa.eu/en/policies/strategy-data.

42. Dans le domaine de la recherche universitaire et des données, le rôle joué par le droit d'auteur, les licences de libre accès et plus généralement les principes de la science ouverte restera central. Cela est dû au moins en partie au fait que le mot « données » a plusieurs définitions, et ce, dans différents domaines. Nous avons essayé de montrer que ce que la science et les sciences humaines et sociales peuvent appeler des « données » ne sont pas nécessairement des données du point de vue du droit d'auteur. S'agissant par exemple des articles de recherche, les chercheurs en sciences humaines peuvent très bien appeler données ce que leur approche disciplinaire et méthodologique identifie comme étant des données. En réalité, l'utilisation des catégories et des processus acceptés et validés dans un champ disciplinaire apparaît comme la seule approche conforme aux principes scientifiques. En d'autres termes, un document de recherche en sciences humaines et sociales devrait appeler « données » les éléments qui, selon les règles des SHS, sont identifiés comme étant des données. Un document de recherche en sciences juridiques peut avoir besoin de suivre une terminologie différente afin de se conformer à ses propres règles de validation scientifique. Le problème peut survenir lorsque ces barrières disciplinaires tombent et que les concepts et les méthodes franchissent les disciplines. Dans ce cas, le développement de méthodes, taxonomies et ontologies communes apparaît comme une première étape dans le développement d'une approche interdisciplinaire appropriée.

43. Une autre difficulté se présentera lorsqu'un chercheur en SHS (ou toute autre discipline) souhaitera savoir si la licence CCPL permet bien la réutilisation de publications ou de données sous-jacentes, la définition de « données » étant différente dans chaque discipline. Pour de telles situations, nous espérons que l'analyse présentée ci-dessous constituera une petite contribution au débat.

Science ouverte, plans de gestion de données et *data papers* au cœur d'une offre de services : l'exemple du SCD de l'université de Lille

Alicia León y Barella

Introduction

1. Le Service commun de documentation (SCD) de l'université de Lille coordonne une ambitieuse politique pour mettre la science ouverte au cœur de son offre de services, dans la continuité des actions initiées par les trois universités (Lille 1, 2 et 3) qui ont précédé la création de l'université de Lille (2018). D'abord centrée autour de la diffusion des publications, via la mise en place de l'archive ouverte institutionnelle LillOA, cette offre de services s'élargit désormais à la gestion, à la diffusion et à la valorisation des données de recherche produites par les équipes de recherche de l'université. Cette offre de services, récemment mise en valeur par la labellisation en « atelier de la donnée », intègre notamment les enjeux liés aux plans de gestion de données, au dépôt des jeux de données et à leur valorisation via les *data papers*.

2. Le SCD comprend quatre équipements documentaires, répartis sur les principaux campus de l'université[1]. Composé d'une équipe de plus de 200 personnes, le SCD assure l'accueil des 78 000 étudiants et des doctorants, chercheurs et enseignants-chercheurs issus des 64 unités de recherche de l'université représentant l'ensemble des disciplines (santé, sciences humaines et sociales, droit et sciences politiques, sciences de l'environnement et technologie). Il met à disposition de la documentation imprimée et électronique, assure des formations pour tous les niveaux, afin de mieux se repérer parmi les outils et services disponibles, et organise régulièrement des animations culturelles. Le SCD travaille en coopération avec d'autres bibliothèques au sein de l'établissement public expérimental (EPE) : bibliothèques associées, bibliothèques de l'ENSAIT, de l'ENSAPL, de Sciences Po Lille, etc.

Le dispositif d'appui à la recherche

3. Le département « Services à la recherche et aux chercheurs » du SCD coordonne et développe l'offre de services, d'outils et de formations spécifiquement dédiée aux chercheurs, doctorants et personnels de recherche, portant notamment sur les bonnes pratiques en matière de science ouverte (accompagnement de projets de recherche, archives ouvertes, données de recherche, aide

1. LILLIAD Learning center Innovation (Campus Cité scientifique), la BU Sciences Humaines et Sociales (Campus Pont-de-Bois), la BU Droit-Gestion (Campus Moulins), la BU Santé (Campus Santé). En savoir plus : https://bu.univ-lille.fr.

à la publication, etc.), mais aussi sur la numérisation de corpus, la diffusion des thèses et mémoires, les identifiants chercheurs, les nouvelles approches bibliographiques, etc.

4. Depuis la réorganisation du SCD à l'issue de la fusion des universités en 2018, un réseau de référents « science ouverte » a été constitué au sein du SCD. Mis en place dans un contexte d'élargissement du périmètre et des missions du SCD, ce dispositif joue un rôle central dans les relations entre le SCD et les unités de recherche. Ces six collègues, répartis sur l'ensemble des campus, ont en charge une dizaine d'unités de recherche chacun et sont les interlocuteurs privilégiés des unités de recherche : ils sensibilisent, forment, accompagnent et orientent les enseignants-chercheurs, doctorants et personnels d'appui à la recherche des unités de recherche sur l'ensemble des sujets liés à la science ouverte.

Le soutien à la science ouverte

5. L'université de Lille porte depuis plusieurs années une politique ambitieuse d'ouverture de la recherche, inscrite dans le mouvement international de la science ouverte. Elle a ainsi créé dès 2018 une archive ouverte institutionnelle, LillOA[2], et mène plusieurs actions pionnières dans le domaine de la gestion des données de recherche. Elle affirme également son engagement en faveur du développement de la science ouverte, à travers une implication volontariste à l'écosystème qui le sous-tend, en soutenant les initiatives vertueuses et en participant à l'effort collectif local, national et européen. Ainsi, le SCD de l'université de Lille a-t-il été choisi pour coordonner la réalisation de la collection *Passeport pour la science ouverte*, projet financé par le ministère de l'Enseignement supérieur et de la Recherche et piloté par le Comité pour la science ouverte[3] visant à sensibiliser les doctorants. Manifestation officielle de l'engagement et des ambitions de l'université, la *Feuille de route pour la science ouverte* de l'université a été publiée à l'été 2021[4]. Elle regroupe des actions pilotées par le SCD (ouverture des publications et des données, soutien à l'écosystème de la science ouverte) et par la Direction Recherche et valorisation (recherches participatives, édition ouverte).

6. Au sein du SCD, les actions et services concernant l'ouverture de la science sont coordonnés par le service science ouverte, qui développe des outils, met en place des dispositifs de sensibilisation, d'accompagnement et de formation sur l'ensemble des volets de la science ouverte. Le service accompagne déjà depuis plusieurs années les équipes de recherche de l'université de Lille, notamment celles qui sont engagées sur des projets financés par l'ANR ou la Commission européenne, pour les aider à se mettre en conformité avec les exigences

2. https://lilloa.univ-lille.fr

3. https://www.ouvrirlascience.fr/passeport-pour-la-science-ouverte

4. https://bu.univ-lille.fr/fileadmin/user_upload/SCD/Recherche/FdRouteSO_A4print.pdf

liées à la science ouverte : diffusion des publications sur une archive ouverte, ouverture des données quand cela est possible, rédaction du plan de gestion de données (PGD). Ce service correspond à un véritable besoin pour les chercheurs, qui ne sont pas encore – pour beaucoup – familiers de ces enjeux. Ainsi, en 2022, nous avons accompagné 82 porteurs de projets différents, soit de façon collective (participation à des ateliers, des formations), soit de façon individuelle (relecture du projet, du PGD, vérification de la conformité des revues choisies pour les publications, etc.).

L'accompagnement pour les données de recherche

7. L'offre de services, d'abord centrée sur les besoins des coordinateurs de projets financés par l'ANR ou la commission européenne, directement concernés par l'obligation de rédiger un plan de gestion de données (PGD), s'est progressivement élargie, avec la prise de conscience de la nécessité de « bien » gérer ses données pour tous les acteurs de la recherche publique. Prenant en compte ce besoin fort des équipes de recherche, le SCD a donc considérablement renforcé son engagement autour de l'aide à la gestion des données de recherche, que ce soit en termes de ressources humaines ou en termes de services, avec par exemple l'ouverture en juillet 2022 de l'espace institutionnel Recherche Data Gouv[5] ou encore la mise en place d'un service d'appui à la FAIRisation des bases de données. La qualité du service apporté a été reconnue via la labellisation en décembre du projet Lille Open research data (LORD)[6] en « atelier de la donnée » dans le cadre de Recherche Data Gouv. Les ateliers de la donnée constituent en effet l'un des piliers de l'écosystème Recherche Data Gouv et visent à apporter aux équipes de recherche un accompagnement de proximité dans la gestion de leurs données de recherche.

8. L'accompagnement proposé aux équipes de recherche de la métropole lilloise prend des formes variées en fonction du niveau de compétence initial et concerne tous les aspects concernant la gestion des données, des codes et des logiciels de la recherche, pour couvrir l'ensemble des besoins des équipes de recherche et de leurs membres : actions de sensibilisation et de formation, accompagnements personnalisés. Cela permet également de simplifier l'identification des différents acteurs de la donnée sur le territoire, de tous les corps de métiers concernés par la gestion des données de recherche, afin de soulager au maximum les équipes de recherche. Les actions et services proposés dans le cadre de l'atelier Lille Open research data sont les suivants :

- **Actions de sensibilisation** : la Fabrique de la science ouverte[7], dispositif de sensibilisation organisé pour expliquer comment la science ouverte se fait, se construit,

5. https://entrepot.recherche.data.gouv.fr/dataverse/univ-lille

6. LORD est un projet porté par l'université de Lille et financé par le Fonds national pour la science ouverte. https://bu.univ-lille.fr/chercheurs-doctorants/projets-et-plateformes/lord.

7. https://fabso.univ-lille.fr

se pratique concrètement dans une recherche. Ce dispositif se construit en synergie avec des actions nationales, comme le *Printemps de la donnée*, ou l'*Open Access week*.

- **Actions de formations collectives** : programme de formation continue pour les personnels d'appui à la recherche et les enseignants-chercheurs, programme de formation doctorale sur les différents volets de la science ouverte (dont les données de recherche), ateliers pour les porteurs de projets financés (phase de montage et phase de suivi des projets), sur les aspects science ouverte.
- **Accompagnement personnalisé des chercheurs** : relecture des plans de gestion de données, aide au montage des projets (relecture du projet sur les aspects science ouverte), conseil et assistance pour tous les sujets liés au cycle de vie des données. Un accompagnement particulier a également été mis en place autour des bases de données réalisées dans le cadre de projets de recherche et de leur mise en conformité avec les principes FAIR. Ainsi les équipes de la plateforme NORINE, base de données de peptides non-ribosomiques, ont-elles bénéficié de cet accompagnement qui leur a permis d'obtenir en 2022 le prix science ouverte des données de la recherche[8].
- **Aide à l'orientation et à l'identification des différents acteurs** : une partie des services mis en place dans le cadre de l'atelier de la donnée (Recherche Data Gouv) vise à faciliter le travail d'identification des différents acteurs de la donnée, répartis entre plusieurs entités et services (unités de recherche, services informatiques, services d'archives, services de documentation, délégué à la protection des données, comité éthique, etc.). Pour ce faire, un Guichet d'assistance unique a été mis en place, qui centralise les différents acteurs de la donnée sur le territoire[9], et un réseau de « correspondants de la donnée » se construit progressivement afin que, dans chaque unité de recherche, un ambassadeur en proximité puisse faire connaître les bonnes pratiques et transmettre aux membres de l'atelier LORD les besoins et questions des équipes.

Le public cible de cet accompagnement

9. Comme on l'a vu plus haut, le public s'est élargi. Visant d'abord les coordinateurs de projets financés, les services proposés par le SCD sont désormais ouverts à l'ensemble des acteurs de l'université de Lille : étudiants, doctorants, enseignants-chercheurs et personnels d'appui à la recherche de l'ensemble de l'EPE. Au vu du caractère pluridisciplinaire de l'université de Lille, nous nous adressons donc à toutes les disciplines scientifiques (sciences humaines, sociales et juridiques, sciences de l'ingénieur, environnement, biologie et santé, etc.).

10. La mise en place de l'atelier de la donnée *Lille Open research data* a permis d'élargir encore ce public initial. En effet, pour répondre aux enjeux de mutualisation à l'échelle du territoire, le projet associe de nombreux établissements

8. Pour en savoir plus sur la FAIRisation de NORINE, une séance de la Fabrique de la science ouverte a été consacrée à ce sujet : https://fabso.univ-lille.fr/nos-evenements/saison-2022/produire-des-donnees-fair-pour-les-rendre-faciles-a-trouver-accessibles-interoperables-et-reutilisables.

9. assistance.univ-lille.fr/formulaire/scienceouverte.

et organismes de recherche du territoire lillois. Outre l'université de Lille, on y trouve des établissements de recherche comme Centrale Lille, l'IMT Nord-Europe, l'ENSAM, le CHU de Lille ou un peu plus loin l'université de Liège, ainsi que des organismes nationaux comme le CNRS, INRIA, INRAE. Par ailleurs, le projet s'est construit en complémentarité avec des « nœuds thématiques » qui viennent apporter une expertise disciplinaire sur la gestion des données produites : la plateforme bilille (membre de l'Institut français de Bioinformatique) et le CHU de Lille pour le domaine Santé et Sciences de la vie ; ICARE (un des quatre centres de données et de ressources d'AERIS, au sein de Data Terra) pour la recherche atmosphérique ; la Maison européenne des sciences de l'homme et de la société (MESHS), liée à Huma-Num et la Plateforme universitaire des données de Lille (PUDL), service de la MESHS, membre de PROGEDO[10], pour les Sciences humaines et sociales ; et enfin l'Office national d'études et de recherches aérospatiales (ONERA) pour l'aérospatial.

Les plans de gestion de données

11. Nous proposons aux coordinateurs de projet de participer à des ateliers collectifs pour mieux comprendre les enjeux autour de la gestion des données, les attendus de la commission européenne et de l'ANR sur le sujet (diffusion, signalement, etc.). La place des Plans de gestion de données y est naturellement centrale : nous revenons en détail sur les rubriques qui composent le Plan de gestion de données, sur les éléments qui peuvent y figurer et que l'on pourra détailler en fonction des spécificités de chaque projet et des particularités de chaque famille disciplinaire. Plus largement sont abordées les différentes étapes du cycle de vie des données, et qui seront, de fait, précisées dans le PGD : comment les données sont-elles collectées, traitées, comment seront-elles diffusées et valorisées ?

12. L'un de nos objectifs est de faciliter la rédaction de ce document, qui peut être perçue comme une nouvelle tâche administrative lourde et inutile. Nous essayons de montrer que la rédaction de ce document peut au contraire être une aide pour anticiper, dès le début du projet, les difficultés méthodologiques, techniques, budgétaires, administratives qui peuvent se présenter en lien avec le traitement des données de recherche.

13. Pour accompagner les chercheurs à rédiger le PGD, nous avons mis en place des outils pratiques et faciles d'utilisation. Ainsi, nous fournissons un modèle de fichier *readme*, sur le modèle de l'université de Cornell, qui permet de décrire le jeu de données pour qu'il soit explicite et donc exploitable par d'autres équipes de recherche. Par ailleurs, nous mettons à disposition un espace institutionnel sur DMP OPIDoR, outil d'aide à la rédaction des plans de gestion de données proposé par l'INIST du CNRS[11]. On trouve sur cet espace un accès facilité

10. Huma-Num et Progedo sont tous deux des centres de référence thématique, Recherche Data Gouv.

11. https://dmp.opidor.fr

à nos services, et prochainement un modèle de PGD propre à l'université de Lille. Basé sur la trame du PGD recommandé par l'ANR, ce document fournira des recommandations et des conseils spécifiquement dédiés à nos équipes de recherche. On y trouvera par exemple un descriptif des outils mis à disposition par l'université pour le partage, le stockage ou la diffusion de ces données, un rappel des personnes-ressources au sein de l'université, etc. L'enjeu pour un document de ce type est d'en faire à la fois un document utilisable et compréhensible, tout en étant adapté pour chaque famille disciplinaire. Ce document sera donc utilement complété par une synthèse des ressources en ligne utiles, qui permettront notamment d'approfondir certains aspects auprès des experts disciplinaires (par exemple Huma-Num, Data Terra, etc.).

14. Enfin, pour accompagner au mieux nos unités de recherche, nous proposons systématiquement une relecture des plans de gestion de données réalisés dans le cadre de projets ANR ou européens. Cette approche nous permet d'adapter nos conseils et outils proposés, en fonction du projet de recherche et des données traitées. Cela permet également d'anticiper avec les chercheurs des questions (enjeux juridiques et éthiques, partage et conservation des données, etc.) qui peuvent avoir été peu ou mal identifiées.

Les *data papers*

15. Lors de nos ateliers dédiés à la gestion des données, nous abordons systématiquement la question des publications qui peuvent valoriser les jeux de données déposés dans un entrepôt, et notamment les *data papers*. Une fois le travail de nettoyage, description et dépôt fait, nous travaillons avec les chercheurs sur leur stratégie de publication en *Open access*, nous en profitons pour évoquer la démarche de rédaction de *data paper*, qui peut permettre une valorisation supplémentaire de la recherche effectuée et des données associées. Il est effectivement dommage de se priver d'un article dans une revue évalué par les pairs et citable.

16. Nous avons de plus en plus de demandes de soutien autour de ce nouveau type de publications afin d'en comprendre le fonctionnement, la nature, et d'obtenir des conseils pour le choix des revues. Nous assistons les chercheurs dans leur rédaction afin de tirer parti du travail déjà réalisé lors du dépôt et de la description des données. Ainsi, les chercheurs peuvent désormais, via l'espace de l'université de Lille dans Recherche Data Gouv, générer automatiquement une première ébauche de *data paper*, à partir du DOI d'un jeu de données publié dans l'entrepôt. Cette ébauche devra ensuite être complétée manuellement, et l'on pourra s'aider de la documentation associée au jeu de données lui-même, notamment le Plan de gestion de données et/ou le fichier Readme. En effet, un certain nombre d'éléments présents dans ces deux documents sont au cœur du

data paper (description, origine, propriété des données, licences et conditions de réutilisation, potentiel de réutilisation). D'autres documents peuvent venir compléter et expliciter le *data paper*, comme un dictionnaire des variables, un protocole d'obtention, etc. Ces documents sont complémentaires et reposent sur le même enjeu : la réutilisation des données rendues publiques.

17. Outre le contenu du *data paper*, l'une des facettes de notre accompagnement repose sur le choix de la revue dans laquelle publier son *data paper*. Pour ce faire, nous pouvons compter sur un certain nombre de ressources précieuses, comme DoraNum[12] ou CoopIST[13]. Par ailleurs, nous réfléchissons – à plus long terme – à des solutions liées à notre plateforme de revues, *Péren*[14], qui donne accès à des revues scientifiques éditées dans les Hauts-de-France. Plusieurs revues comme *Lexique*[15] ou *Espace, Populations, Sociétés*[16] sont d'ores et déjà intéressées par le dépôt des données sur l'espace institutionnel de Recherche Data Gouv et nous leur avons donc créé des espaces propres afin que les auteurs qui publient dans ces revues puissent également déposer les jeux de données associés à leurs articles. Dans la continuité de ce service, la question se pose de la publication sur *Péren* de *data papers* associés aux jeux de données déposés sur Recherche Data Gouv, afin de valoriser ces dépôts, par exemple via un data journal pluridisciplinaire. Ce nouveau développement permettrait de créer des passerelles entre différents types de productions de l'université et entre différents projets liés à la science ouverte. Ces projets offrent de belles perspectives pour l'avenir, pour la valorisation, la diffusion et la réutilisation des jeux de données rendus publics par la communauté de recherche.

12. https://doranum.fr/data-paper-data-journal
13. https://coop-ist.cirad.fr/gerer-des-donnees/publier-un-data-paper
14. Péren (Plateforme d'édition de revues numériques) est un projet porté par l'Université de Lille, en partenariat avec la MeSHS et OpenEdition, et financé par le fonds national pour la science ouverte. La plateforme donne ou donnera accès à plusieurs revues éditées ou soutenues par l'université de Lille. https://www.peren-revues.fr/.
15. https://lexique.univ-lille.fr/
16. https://journals.openedition.org/eps/

Des corpus de langage oral aux *data papers*

Christophe Parisse

Introduction[1]

1. Ce chapitre a pour but de décrire l'état de l'art dans la structuration et la distribution de corpus de langage oral et de montrer en quoi la diffusion de ces corpus partage beaucoup de points communs avec la diffusion des *data papers* et des données qui leur sont associés. Dans le reste du chapitre, on présentera des solutions pour passer d'un corpus à un authentique *data paper* avec ses données, qui puisse être publié dans des revues scientifiques avec comité de lecture, ainsi que les avantages que cela peut offrir à la recherche dans le domaine du langage oral.

Qu'est-ce qu'un *data paper*, qu'est-ce qu'un corpus oral ?

2. Pour comparer corpus et *data paper*, il faut présenter ce que sont les *data papers*, même si cette présentation pourra être largement trouvée dans d'autres chapitres de cet ouvrage, avec probablement des petites différences d'un domaine scientifique à l'autre. Le *data paper* est une publication scientifique comprenant essentiellement une description de données de recherche accessibles qui peut s'accompagner de nombreuses informations scientifiques, mais sans comporter de démonstration scientifique.

3. Dans une publication traditionnelle, on ne trouve que rarement des détails complets sur les données de recherche qui ont servi à sa réalisation (même si cela est fortement recommandé dans certains domaines scientifiques ou dans certaines revues, voir par exemple Obels *et al.*, 2020). On trouve plutôt (pour les meilleures publications) les éléments permettant la reproduction de la recherche et les éléments théoriques à démontrer associés à une discussion des résultats. Dans un *data paper*, on trouvera une description précise et exhaustive, la plus informée possible, des données de telle manière qu'elles puissent être contrôlées ou utilisées pour un travail scientifique. Ces données doivent être librement accessibles et les conditions juridiques et techniques de leur réutilisation doivent être claires.

4. Si un *data paper* ne contient pas, par définition, les données qu'il décrit, il est toutefois indissociable de ces données sans lesquelles il n'a pas de valeur. Inversement, la valeur des données est enrichie de manière multiple par l'existence d'un *data paper*. Des données sans description n'ont que peu de valeur. Dans ce chapitre, je parlerai toujours du couple *data paper* et données, parce que c'est avec ce

1. De grands remerciements à Aliyah Morgenstern et Aude Da Cruz Lima pour leurs remarques, corrections et suggestions sur une première version de ce chapitre.

couple que je travaille en tant que scientifique, et parce que c'est ce couple que je peux comparer avec le couple formé par les corpus de langage et leurs descriptions.

5. Un corpus de langage (Dalbera, 2002) oral est un ensemble de données issues d'un recueil de langage parlé (en opposition au langage écrit)[2]. Si cet ensemble de données peut exister sans être décrit, une telle situation rendrait le corpus peu utilisable et peu intéressant. C'est pourquoi le corpus sera très souvent accompagné de descriptions et de publications qui, même si elles n'en ont pas le label, ressemblent à des *data papers*. Si le parcours des corpus va souvent des données à la description, le couple que forment le corpus et sa description ressemble au couple que forment le *data paper* et ses données.

6. Si la notion de *data paper* est assez récente (voir par exemple Chavan et Penev, 2011), les corpus de langage, y compris de langage oral, sont bien que plus anciens (voir ci-dessous, Kucera, 1967, pour le premier exemple). Ils suivent depuis très longtemps les mêmes principes que les *data papers* et leurs données, même si on peut trouver des différences. C'est pourquoi la comparaison historique de ces deux objets scientifiques est intéressante. Elle peut aider à faire comprendre la grande qualité des corpus de langage oral utilisés dans la recherche, et aussi permettre de savoir ce qu'il peut rester à faire pour faire passer les corpus de langage oral de données scientifiques partagées et leur description à de véritables *data papers* accompagnés de données, et en quoi ce passage peut être important pour la linguistique.

Les corpus de langage oral : historique et spécificités

7. Le recueil de données de corpus oral peut être sonore (avec un magnétophone) ou audio/visuel (avec une ou plusieurs caméras). Les données recueillies sont le plus souvent accompagnées de transcriptions (description du langage parlé en utilisant une version adaptée du langage écrit) ou/et de codages (ensemble d'analyses qui caractérisent les données recueillies). Les corpus de langage oral, lorsqu'ils sont déposés et accessibles sur un site internet, forment un type de données dont les propriétés intrinsèques en font un ancêtre naturel des données FAIR. Ces corpus sont le plus souvent associés à des articles ou ouvrages scientifiques les utilisant. Ces articles ressemblent donc à des *data papers*, c'est-à-dire des publications scientifiques dont l'objet est la description de données qui seront utilisées pour d'autres recherches que celles des auteurs des données.

8. Si les corpus de langage oral ne sont pas systématiquement organisés pour être décrits par des revues spécialisées dans ce domaine, ils présentent néanmoins

2. Dans ce chapitre, la notion de corpus de langage oral est mise en avant. De fait, la plupart des caractéristiques des corpus oraux se retrouvent dans les corpus écrits. La seule différence entre les deux types de corpus (du point de vue de la comparaison avec les *data paper* et leur données), est que le coût humain et pécuniaire des corpus oraux rend encore plus critique la question du dépôt des données et de leur description. Cette question s'applique toutefois aussi à de nombreux exemples de corpus écrits, qui peuvent également être très coûteux et précieux.

plusieurs caractéristiques fondamentales auxquelles ils ne peuvent échapper :

1. Ils sont coûteux à produire, en temps et/ou en argent.
2. Ils sont rares et donc ils ont tendance à être les données de référence dans les travaux scientifiques.
3. Ils sont réutilisables de nombreuses fois pour des recherches très variées.
4. Ils représentent souvent l'investissement total d'une ou plusieurs personnes, pendant plusieurs années, et justifient d'être reconnus comme des productions scientifiques de valeur.

9. Ces propriétés se retrouvent dans d'autres domaines que celui des corpus de langage oral (par exemple les corpus de langage écrit). Elles sont peut-être ici un peu plus saillantes, plus lourdes de sens et coûteuses pour les personnels de recherche et pour les institutions, en raison du travail manuel encore important à fournir pour le recueil et la préparation des données de l'oral. On remarquera que l'utilisation, dans des corpus récents, de données de langage plus importantes, plus automatisées, grâce en particulier aux traitements automatiques non supervisés, ne change pas l'importance du coût des données et de leur description scientifique. On passe de données analysées finement en détail pour des petits corpus à des données beaucoup plus volumineuses et souvent analysées avec moins de détail. Mais d'un point de vue scientifique, ces données doivent être informées et décrites en détail. Ceci est vrai, quel que soit l'usage de ces données, à la fois pour l'analyse manuelle de certaines données précises recherchées dans un grand ensemble ou pour le traitement massif en vue de traitements automatiques.

10. Les corpus de langage oral déposés et accessibles sont donc pour leur totalité conçus pour une réutilisation des données et accompagnés de description, y compris dans des articles scientifiques publiés dans de grandes revues scientifiques. Néanmoins, cette réutilisation n'en fait pas nécessairement des *data papers* accompagnés de leurs données. Certes, presque tous les corpus connus sont utilisables par des chercheurs n'ayant pas participé à leur recueil. Mais, dans certains cas, cette utilisation est limitée, soit volontairement pour contraindre l'utilisation du corpus et ne pas l'offrir librement, soit simplement parce que la technologie utilisée pour mettre le corpus à disposition n'est pas ouverte ou conçue pour le partage des données.

11. Les différences que l'on peut trouver entre corpus de langage et leur description et les *data papers* et leurs données portent d'abord sur le côté systématique d'une validation par un comité de lecture scientifique, comme pour les *data papers*. Le fait que les corpus de langage oral aient le plus souvent mené à une publication scientifique va dans cette direction. Une autre différence pour les corpus concerne les formats, les dépôts, et d'une manière générale, le respect des principes FAIR[3] (voir ci-dessous « Propriétés des corpus de langage oral »). Une difficulté rencontrée dans le partage des corpus est celle du choix

3. https://www.go-fair.org/fair-principles/

de technologies, de normes et de formats facilement partageables, connus et bien décrits. Il s'agit ensuite de respecter ces choix, qui nécessitent parfois de grands investissements, en particulier lorsque l'anonymisation des corpus est une nécessité. L'anonymisation implique de cacher toute information qui permettrait d'identifier les personnes qui parlent et, par là même, permet de faire sortir les données du domaine privé, c'est-à-dire d'en faire des données publiques, non personnelles, qui ne sont pas contraintes par le RGPD. Si les données ne peuvent être rendues anonymes, alors leur diffusion ne peut se faire qu'avec le consentement éclairé de toutes les personnes figurant dans les corpus.

Les corpus de langage oral en tant que données de recherche

Historique

12. Il existe quelques corpus historiques dont la création date d'avant 1990 comme le Brown corpus (Kucera et Francis, 1967) ou le corpus Cobuild (Renouf, 1984), mais l'un des plus connus et représentatifs corpus de langage est le British National Corpus (BNC) dont l'historique est décrit par Burnard (2002). Ce corpus n'est pas le seul des grands corpus anglophones et on peut aussi citer l'Open American National Corpus (OANC, voir Ide, 2013) dont l'objectif initial était de reproduire le BNC en anglais américain.

British National Corpus (BNC)

13. Le BNC (Aston et Burnard, 2020), corpus d'environ 100 millions de mots, a deux particularités qui sont intéressantes pour faire le lien technique et historique avec les *data papers* et leurs données. D'une part, les données ont été rendues publiques et sont disponibles librement pour la recherche, avec une licence ouverte qui définit les droits d'utilisation du corpus[4] et avec un identifiant pérenne. D'autre part, le développement du BNC a conduit à définir des formats et des normes pour la présentation de corpus. Ce développement a suivi de manière parallèle (et souvent avec les mêmes acteurs) celui de la *Text Encoding Initiative* (TEI). La TEI[5] est une référence pour le codage de textes langagiers et de corpus. La TEI est dirigée par une large association de laboratoires et de chercheurs, ce qui en fait un outil très partagé et consensuel, même si, bien sûr, il ne peut être parfait. La conséquence est surtout que, comme les textes du BNC sont codés suivant les principes de la TEI, ils utilisent de fait un format ouvert permettant une réutilisation aisée des corpus, y compris de manière automatique.

14. Le BNC contient environ 10 % de contenu correspondant à de l'oral, comprenant des conversations et des enregistrements réalisés dans des situations plus formelles. Malheureusement, le BNC ne contient pas les enregistrements sonores, mais simplement les transcriptions,

4. BNC User Licence www.natcorp.ox.ac.uk/docs/licence.html.
5. http://www.tei-c.org/

ceci pour des raisons historiques, car, à l'origine du projet, stocker et distribuer les sons paraissait trop lourd techniquement pour être réalisé. Les données orales comportent des informations sur les locuteurs et sur la situation de production sous forme de métadonnées insérées dans les fichiers au format TEI, ce qui fait du BNC un bon exemple de corpus permettant la réutilisation scientifique automatique.

Open American National Corpus (OANC)

15. La situation du BNC n'est pas unique, car l'OANC a des caractéristiques proches, mais différentes, ce qui montre que la notion de *données partagées et décrites* est facilement mise en œuvre, mais pas toujours réalisée de la même manière. L'OANC n'a pas bénéficié de financements importants, comme le BNC et a donc eu une histoire moins proche des institutions. Ceci explique peut-être que le corpus, bien qu'accessible très aisément sur le site web de l'ANC[6], ne soit pas protégé par un identifiant pérenne dans un dépôt officiel. Également, le choix de format du corpus a été effectué de manière différente. Le choix s'est porté sur le format GrAF/LAF, qui est une norme ISO 24612[7] qui est plus utilisée par la communauté du traitement automatique des langues que celui des humanités numériques et de la littérature (comme l'est la TEI).

CHILDES/TalkBank

16. Un dernier exemple historique de corpus qui peut servir d'exemple pour la diffusion de données de corpus de langage oral associées à leur description est celui du projet CHILDES, étendu par la suite au projet TalkBank[8]. CHILDES est remarquable par son volume, par son organisation, et par le fait qu'il contient des exemples qui pourraient former un couple données et *data paper*. Son origine date de 1981, mais sa présentation officielle a été faite dans un article de MacWhinney et Snow (1985) qui explique l'intérêt de créer une base de données, de définir des formats, de créer des logiciels pour les utiliser. De manière très intéressante, les auteurs citent trois raisons majeures pour justifier la création de leur programme, qui sont toutes également d'actualité pour les *data papers* et leurs données :

- Le besoin d'une plus grande efficacité dans le partage des données (*The need for greater efficiency in data sharing*, p. 273)
- Le besoin de plus de précision dans la collecte et l'analyse des données (*The need for greater precision in data collection and analysis*, p. 273)
- Le besoin d'améliorer l'automatisation des analyses (*The need for increased automation in analysis*, p. 274)

17. Une description plus complète est donnée dans MacWhinney et Snow (1990) qui présente les trois points qui ont fait la réussite du projet :

6. https://www.anc.org/
7. Cf. https://www.iso.org/obp/ui/#iso:std:iso:24612.
8. https://www.talkbank.org/

- CHILDES : La base de données ouvertes de corpus de langage
- CHAT : Les conventions de codage des transcriptions.
- CLAN : Un outil permettant d'éditer les transcriptions.

18. Ces trois points forment un triptyque incontournable de la réussite des principes FAIR.

1. CHILDES est une base de données ouverte et accessible. Elle peut être moissonnée automatiquement et il est donc possible de récupérer ces données dans des moteurs de recherche comme celui d'Olac[9] ou du VLO de CLARIN[10]. Ces moteurs de recherche sont, à la différence de ceux de Google ou de Microsoft, spécialisés dans la recherche d'informations de corpus.
2. CHAT est un système de convention pour le formatage de transcriptions de données orales, comprenant des instructions pour le codage des métadonnées comme pour celui des données. Il est ainsi possible de savoir comme manipuler automatiquement les données de la base CHILDES. Ceci permet l'interopérabilité, puisqu'il est possible de créer des conversions automatiques des données.
3. CLAN est un outil informatique permettant d'éditer, visualiser et jouer les données de langage oral. L'existence de l'outil est fondamentale, car elle garantit la qualité des données, mais aussi leur réutilisabilité.

19. CHILDES est particulièrement connu pour une dernière raison qui est celle d'avoir largement incité à respecter la citation des données, si importante pour les chercheurs modernes. Des règles d'usage de la base de données ont été instaurées depuis sa création[11] et ont amené à citer largement les corpus déposés dans la base. L'usage classique dans CHILDES est celui de citer la base de données en général et surtout un des articles produits par les auteurs de corpus, comme indiqué dans le dépôt du corpus lui-même. Il est aussi possible de citer uniquement le dépôt par l'utilisation d'un DOI (Digital Object Identifier : un identifiant informatique pérenne, donc unique et permanent) comme indiqué dans la page « citation.html ». Par exemple : « Bat-El, Outi (2020). CHILDES Database Hebrew Bat-El Corpus. DOI : 10.21415/T5P322 », fait référence à un corpus qui n'est pas associé à un article scientifique. Il s'agit donc exactement d'une donnée de recherche qui n'est pas passée par le comité de lecture d'une revue spécialisée. L'intérêt de tels dépôts est clair. Si l'on prend l'exemple d'un des tout premiers dépôts de corpus en langue française réalisé sur CHILDES, celui de Christian Champaud 1994 (CHILDES French Champaud Corpus, DOI : 10.21415/T5M88F)[12], on constatera que le contenu de la citation (Champaud, 1994) est très difficile, sinon impossible à trouver aujourd'hui, car il s'agit d'une présentation dans un congrès, tandis que le corpus en tant que donnée ouverte à la recherche, pouvant être citée, est toujours accessible.

9. http://dla.library.upenn.edu/dla/olac/index.html
10. https://vlo.clarin.eu
11. Cf. https://talkbank.org/share/citation.html.
12. Visible ici https://childes.talkbank.org/access/French/Champaud.html.

20. Plus récemment, les principes de CHILDES ont été étendus à TalkBank dont une description rapide, mais complète, peut être trouvée dans la page de description du TalkBank Clarin Knowledge Centre[13]. Cette dernière version, plus complète et surtout plus avancée technologiquement, conserve les principes qui font de ces données un ensemble préfigurant ce qui pourrait être décrit dans des *data papers* en linguistique de l'oral.

La situation en 2024

21. En 2024, il n'existe pas encore de moyen bien organisé de publier des données de corpus en tant que donnée accompagnée de *data paper*. Par contre, il existe depuis plus de 15 ans des sites de dépôt dans lesquelles il est possible de déposer de manière ouverte des données de corpus. Ces dépôts s'accompagnent d'un identifiant pérenne. Plus exactement, deux centres ont été créés en 2006 dans le cadre du programme du Centre de Ressources sur la Description de l'Oral (CRDO) et validés par le CNRS comme centres de ressources numériques (CRN). Les deux centres existent toujours avec des objectifs et des pratiques légèrement différentes entre eux.

22. Le premier centre, créé à l'origine par le LPL (UMR CNRS Aix-en-Provence), a été d'abord nommé SLDR pour être intégré ensuite dans un projet plus ambitieux ORTOLANG, porté par six laboratoires de recherche et qui a ajouté au stockage des données orales celui des données écrites. ORTOLANG est aujourd'hui un centre permettant de recevoir tous les types de corpus de langage à l'adresse de son site[14]. Le dépôt est simplifié pour permettre un accès facile. Une validation des données est réalisée pour contrôler les métadonnées fondamentales permettant l'accès aux données et la définition du droit des données (licence). Le fonctionnement du dépôt est assez proche de celui d'une archive ouverte. Les formats acceptés sont variés et on trouve des corpus de langage oral, de langage écrit, ou multimodaux ainsi que des dépôts de lexique, de thésaurus et d'outils.

23. Le second centre, géré aujourd'hui par trois laboratoires, permet également le dépôt de tout type de corpus, mais avec un travail documentaire fin et une orientation plus spécifique autour des corpus de langage oral, comme l'indique le nom actuel du site de dépôt, Collections de Corpus Oraux Numériques (COCOON)[15]. On peut par exemple trouver dans ce dépôt la collection Pangloss[16], représentative des travaux de la recherche française sur la conservation des langues rares et peu décrites.

24. Pour les deux centres ci-dessus, tous les outils existent pour déposer et créer par là même des objets qui pourront être pointés par des *data papers*, ou ressembleront au couple *data paper* données si leur dépôt est accompagné

13. https://www.clarin.eu/blog/talkbank-clarin-knowledge-centre

14. Cf. www.ortolang.fr.

15. https://cocoon.huma-num.fr

16. https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-af3bdofd-2b33-3bob-a6f1-49a7fc551eb1

de description scientifique, de préférence publiée dans un journal scientifique (ce qui est par exemple encouragé, mais pas contraint dans le site de dépôt ORTOLANG). Soulignons qu'il existe d'autres outils de dépôt et de pérennisation de données, notamment l'outil Nakala fourni par Huma-Num[17]. Toutefois, ces sites sont à distinguer des journaux de *data papers*, même si, pour soumettre un *data paper*, il convient de respecter un ensemble de contraintes similaires pour les données et les métadonnées, notamment les principes FAIR, mais pas seulement. Ces contraintes, qui se sont clarifiées ces dernières années et qui ont été largement discutées par la communauté des linguistes et notamment le consortium Huma-Num CORLI[18], sont présentées et commentées dans la suite de ce chapitre.

Faire d'un corpus un *data paper* accompagné de ses données : les enjeux

Quelles sont les particularités d'un *data paper*, des métadonnées et des données qui lui sont associées ?

25. Le *data paper* est d'une certaine façon agnostique par rapport aux résultats de recherche que les données pourront permettre d'obtenir. C'est la qualité des données qu'il décrit qui lui donne sa valeur et il a pour objectif de permettre la réutilisation de ces données pour des recherches, quelles qu'elles soient.

26. Dans un *data paper*, on décrit non seulement les données, mais aussi les moyens, les techniques, et les conditions qui ont permis de les obtenir. Ainsi, trois éléments sont systématiquement présents dans un *data paper* : une introduction, une description des données et la méthodologie avec l'équipement nécessaire pour la reproduction des données. Ceci est extrêmement important, car, sans ces éléments, la reproduction du travail scientifique n'est pas possible ou très difficile.

27. Un *data paper* contient donc les métadonnées qui sont toutes les informations annexes permettant d'accéder, de connaître, d'utiliser les données mises en partage. Les formats des métadonnées doivent suivre des standards très précis pour qu'elles puissent être automatiquement retrouvées et analysées. Les formats des données présentent plus de latitude, car ces formats peuvent être indiqués dans les métadonnées, alors que les métadonnées doivent respecter les formats des bases de données dans lesquelles elles figurent. Il est donc possible de choisir pour les données les formats et d'autres informations importantes à condition de mettre toutes les explications dans les métadonnées. Toutefois, le choix des formats doit respecter certaines conditions. En particulier, il doit être connu de la discipline, ouvert (c'est-à-dire accessible à tous, sans contrainte), aisé à manipuler automatiquement. Toutes ces contraintes font

17. https://nakala.fr/
18. CORLI (Corpus, Langues, et Interaction) est un consortium labellisé par l'IR Huma-Num (https://www.huma-num.fr/les-consortiums-hn/), cf. https://corli.huma-num.fr/.

partie de ce qu'on appelle les conditions FAIR. Elles ont été largement discutées et validées par la communauté et font aussi partie des réflexions du consortium CORLI. Elles sont présentées dans le cadre des corpus de langage oral ci-dessous, mais peuvent s'appliquer à tous types de métadonnées et de données.

Propriétés des corpus de langage oral Fair

28. Les corpus de langage oral qui sont rendus accessibles pour de futures réutilisations par la recherche doivent présenter toutes les propriétés des objets dits FAIR[19]. FAIR vient de l'anglais : **F**indable, **A**ccessible, **I**nteroperable, **R**eusable. Traduits en français, ces principes comprennent :

- Être **F**acile à trouver : les objets doivent être renseignés avec des métadonnées complètes, claires et explicites. Ils doivent être accessibles via un identifiant pérenne et enfin être référencés sur des sites connus et facilement disponibles.
- Être **A**ccessible : on doit pouvoir retrouver métadonnées et données facilement, avec des protocoles connus et ouverts. Toutes les informations d'accès doivent être clairement indiquées et pouvoir être appliquées automatiquement de manière non ambiguë.
- Être **I**nteropérable : les données doivent pouvoir être utilisées avec d'autres applications, pour d'autres usages. Elles doivent donc être dans des formats connus, bien identifiés, ouverts et utilisables.
- Être **R**éutilisable : les données et métadonnées doivent être richement décrites afin de pouvoir être répliquées ou combinées avec d'autres usages. En effet, le format (aspect interopérable) ne suffit pas toujours à comprendre tout le détail des données, et donc, une description supplémentaire, des exemples d'utilisation, doivent être fournis. Enfin, il faut que la licence et les droits d'utilisation des données soient indiqués, et bien sûr, qu'il y ait moyen d'utiliser les données (ce qui n'interdit pas de contraindre l'accès pour des raisons déontologiques, par exemple).

29. Remplir les conditions « Facile à trouver » et « Accessible » est souvent assez simple, car les sites permettant de déposer, conserver et diffuser des corpus sont en général organisés pour imposer le respect de ces conditions à l'utilisateur. Il faut toutefois faire attention à ne pas confondre des sites de dépôts comme COCOON ou ORTOLANG, avec des sites de sauvegarde des données comme Dropbox, Google Drive ou Microsoft. Un site de dépôt s'accompagne de la création d'identifiant pérenne et rendra disponible les métadonnées dans des bases de données que les moteurs de recherche spécifiques à la recherche scientifique comme Isidore d'Huma-Num[20], le VLO (Observatoire de langage virtuel) de Clarin[21], ou Olac[22] sauront moissonner, c'est-à-dire traiter pour les rendre accessibles à la recherche de données et à la recherche scientifique.

19. Cf. https://www.go-fair.org/fair-principles/.

20. https://isidore.science/

21. https://vlo.clarin.eu

22. http://dla.library.upenn.edu/dla/olac/index.html

30. Plus précisément, ces métadonnées doivent contenir toutes les informations de localisation, de propriété, de droits d'utilisation, de format, de date, des données. Ces informations figurent dans les champs de métadonnées les plus classiques, comme indiqué par exemple dans le Dublin Core[23]. Il est possible de définir des métadonnées d'une plus grande finesse, par exemple en utilisant les recommandations de Olac[24], soit en utilisant les sites de dépôt de corpus comme ORTOLANG ou COCOON, soit en utilisant l'outil créé par CORLI, Teimeta[25]. Le format des métadonnées elles-mêmes peut suivre plusieurs modèles, souvent selon les sites de dépôts. Tous ces modèles peuvent être convertis entre eux automatiquement.

31. Les conditions « Interopérable » et « Réutilisable » sont plus difficiles à remplir, car il existe très peu de sites de dépôt qui vont complètement guider l'utilisateur afin de parfaitement remplir ces informations. La raison en est qu'il existe de nombreuses formes de corpus qui peuvent être interopérables et de nombreux formats. De plus, les standards sont susceptibles d'évoluer à l'avenir avec l'évolution des outils et des technologies. Éviter ces limites peut se faire en utilisant des formats libres.

32. De fait, ces limites sur l'interopérabilité et les formats se rencontrent moins avec les corpus de langage oral car ces corpus présentent des propriétés qui contraignent la variété des formats qu'ils peuvent avoir. La raison en est que les corpus de langage oral sont souvent accompagnés de données multimédia, auditives ou visuelles. Ces données sont complexes à manier, il ne suffit pas d'ouvrir un simple traitement de texte, ce que l'on peut se contenter de faire lorsque l'on travaille, par exemple, sur des corpus écrits. De fait, il est nécessaire d'utiliser certains logiciels qui guident très strictement l'utilisateur. Le nombre de ces logiciels n'est pas très important, car ils présentent un important coût de développement sans faire partie d'un domaine très développé dans l'industrie privée du logiciel. Ainsi, les logiciels les plus utilisés en France sont Praat (spécialisé dans l'analyse phonétique)[26], Clan (spécialisé dans l'acquisition du langage et l'analyse conversationnelle)[27], Elan (spécialisé dans les langues rares et les langues signées)[28] et Transcriber (spécialisé dans la transcription rapide de longs corpus)[29]. Tous ces outils produisent des fichiers parfaitement organisés et qui peuvent être réutilisés automatiquement avec les logiciels qui les ont créés ou avec d'autres logiciels qui reconnaissent ces formats. L'organisation produite par les logiciels de transcription est plus rigide que celle d'un logiciel de traitement de texte, ce qui rend plus aisée leur utilisation automatique.

23. https://www.dublincore.org
24. Cf. http://www.language-archives.org/documents.html.
25. Cf. https://ct3.ortolang.fr/teimeta/ et https://ct3.ortolang.fr/teimeta-doc/.
26. https://www.fon.hum.uva.nl/praat/
27. https://dali.talkbank.org/clan/
28. https://archive.mpi.nl/tla/elan
29. http://trans.sourceforge–net

33. Cette situation a amené le consortium Huma-Num IRCOM[30] puis son successeur CORLI[31] à proposer un outil de conversion des fichiers produits par les logiciels d'annotation de l'oral vers le format défini par la Text Encoding Initiative[32]. Ce format bénéficie de plus d'une norme ISO (ISO 24624:2016) pour les corpus de langage oral. Les conversions peuvent se faire non seulement depuis Praat, Clan, Elan et Transcriber, mais aussi depuis un fichier texte ou au format Word Open Office[33].

34. Les corpus de langage oral qui n'utilisent pas de tels logiciels, soit parce qu'ils ne proposent que du texte libre sans média associé, soit parce que l'association texte/média est une simple association de fichier et non un alignement avec indication de frontières temporelles, sont plus difficiles à intégrer dans un dépôt FAIR. Dans ce cas de figure, il est conseillé, comme c'est le cas pour les corpus écrits, de passer par un format standard connu comme celui de la TEI. L'outil de conversion de CORLI est une des façons de passer au format TEI dans ce cas.

30. http://ircom.huma-num.fr
31. https://corli.huma-num.fr/
32. https://tei-c.org
33. Cf. https://ct3.ortolang.fr/teiconvert/.

Transformer un corpus de langage oral et sa description en un *data paper* accompagné de ses données

35. Le dépôt d'un corpus « FAIRisé », s'il est accompagné d'un article scientifique, en fait un ensemble donnée/description parfaitement adapté à la réutilisation pour la recherche, qui peut être utilisé et cité par les pairs. Par exemple, l'article de Morgenstern et Parisse (2012), « The Paris Corpus » dans *Journal of French Language* contient la description des conditions de recueil, de création et de dépôt du corpus « Colaje » déposé sur le site ORTOLANG[34]. De ce fait, la citation de cet article a pour principal intérêt le respect des conditions d'utilisation du corpus déposé par Aliyah Morgenstern et moi-même sur ORTOLANG. Ce respect indique que les personnes utilisant notre corpus doivent citer notre article. Effectivement, on trouve, au 16 mars 2022, 91 citations de cet article sur Google Scholar[35], un site spécialisé dans la recherche de citations d'articles scientifiques. On constate donc que notre corpus, utilisé par d'autres chercheurs, a quasiment le statut et les propriétés du couple formé par un *data paper* et ses données. Ses citations portent essentiellement sur les données et la méthodologie, et non sur les recherches scientifiques que nous avons faites avec ce corpus. De plus, comme l'article lui-même a été validé par les pairs, car il a été publié dans une revue

34. https://hdl.handle.net/11403/colaje/v2.4
35. https://scholar.google.com/

scientifique avec évaluation par les pairs, il manque à ce travail essentiellement une enveloppe générale qui regrouperait toutes les données, les métadonnées, les articles, les descriptions sur un même support pour en faire un *data paper* accompagné de ses données.

36. À partir de notre exemple, on peut comprendre ce qu'il est nécessaire de faire pour passer d'un dépôt de corpus à un couple *data paper* et données. Tout d'abord, il faut déposer à côté du corpus et des métadonnées des informations les plus précises possibles sur la méthodologie, la description des données, les outils liés aux données. Tous ces éléments permettront à un utilisateur novice de parfaitement comprendre comment le travail a été effectué, de le reproduire le cas échéant, et d'utiliser au mieux les données dans de futurs travaux. Le format à donner à ces éléments n'est pas encore parfaitement normalisé, mais pourra le devenir lorsque davantage de revues contenant des *data papers* pour les corpus de langage oral existeront. Un bon repère est de faire la même description que pour la méthodologie d'un article classique. Il est toutefois souvent possible, pour un *data paper*, d'aller plus loin dans le détail, là où des revues classiques refuseront la publication, car l'ajout d'informations détaillées amène à un dépassement de la taille limite pour un article classique. Enfin, il est toujours utile de rappeler si possible le cadre théorique qui a amené à la collecte de données. En effet, cette collecte est souvent dépendante des présupposés de la recherche, et seul le cadre théorique permet de clarifier ces présupposés.

37. Une dernière contrainte pour réaliser un *data paper* parfait est d'avoir une évaluation par les pairs. C'est possible en passant par une revue classique dans laquelle les données sont décrites, souvent à l'occasion d'une publication de recherche traditionnelle. C'est le plus souvent cette pratique que l'on trouve dans le site CHILDES. L'idéal est bien sûr de passer par une véritable revue de *data papers*. En effet, dans ce cas, non seulement l'évaluation sera faite, mais elle sera faite avec l'objectif de contrôler si le format du dépôt correspond aux conditions des *data papers*.

Déposer un *data paper* en linguistique en 2024 ?

38. Il n'existe pas encore aujourd'hui de revue consacrée aux *data papers* en linguistique. Ceci veut dire que la procédure de dépôt d'un corpus en tant que *data paper* devra se faire manuellement en profitant des services offerts par des sites ayant une fonction plus générale. Le dépôt en lui-même ne pose pas de problème majeur, car plusieurs outils existent aujourd'hui. ORTOLANG et COCOON ont déjà été largement cités dans ce chapitre, et ce sont les sites à privilégier, car ils sont libres, soutenus par la recherche française publique, gratuits, et ils s'engagent à respecter au moins les deux premiers aspects de la recherche FAIR : Facile à trouver et Accessible. D'autres sites peuvent être envisagés pour un dépôt. Comme des sites de dépôt plus génériques, comme celui de Nakala, ou des sites de dépôt et diffusion de données universitaires. Des sites à l'étranger sont aussi

utilisables, comme celui de CHILDES. On pourra aussi utiliser des sites de dépôt pour la science ouverte comme l'Open Science Framework[36] ou Zenodo[37]. Ces sites ont l'avantage de la visibilité internationale, mais l'inconvénient que souvent les dépôts doivent être de petite taille, à l'inverse des sites spécialisés dans la linguistique de corpus. Des dépôts de taille réduite peuvent aussi passer par des sites de dépôts d'articles scientifiques comme HAL-SHS.

39. Dans tous les cas ci-dessus, il faudra toutefois gérer soi-même les deux derniers aspects de la recherche FAIR : interopérable et réutilisable. Pour cela, il faudra utiliser, pour les corpus eux-mêmes, des formats ouverts, non propriétaires, et largement répandus. Il faudra accompagner le dépôt du plus grand nombre de renseignements possible pour assurer la réutilisation des données. Enfin, comme il n'existe pas encore de revues publiant des *data papers* en linguistique orale ni de validation du dépôt par un comité d'évaluation scientifique, on peut suggérer d'inclure dans le dépôt une recherche (projet ou publication scientifique) utilisant les données.

36. https://osf.io/

37. https://zenodo.org/

L'avenir des corpus de langage oral

Le développement des outils de publication

40. Il est relativement facile en 2024 de trouver des corpus de langage. Les corpus de langage oral sont un peu plus rares, car ils sont plus coûteux à produire et à proposer, mais on les trouvera aussi, soit sur des sites de dépôt comme ORTOLANG, COCOON, TalkBank, soit dans des moteurs de recherche spécialisés comme Isidore, le VLO, OLAC (voir ci-dessus partie « Facile à trouver »). Par contre, dans tous ces sites, les corpus, même s'ils sont complètement FAIR, ne seront pas exactement des *data papers* accompagnés de données, puisqu'il leur manque l'évaluation scientifique.

41. Cette situation devrait probablement changer dans le futur, grâce au développement de la recherche ouverte. Ainsi, on peut voir que, pour les recherches en linguistique computationnelle (traitement automatique du langage par ordinateur), les données et les programmes sont de plus en plus disponibles (cf. Wieling *et al.*, 2018). Le partage des données est plus avancé dans le traitement automatique du langage, mais on trouve également de plus en plus de voix en linguistique classique pour défendre cette approche (cf. Berez-Kroeker *et al.*, 2018). On peut donc s'attendre à ce que le partage des données devienne la norme et que la notion de *data papers* en linguistique se développe.

42. Cette tendance devrait se renforcer avec l'exigence de constitution de Data Management Plan (DMP) dans le cas de projets obtenus avec des fonds publics comme l'ANR (Agence Nationale de la Recherche). Les DMP sont des outils qui permettent de décrire finement les données et d'aider les scientifiques à conserver, sécuriser et diffuser leurs données[38].

43. Il n'existe pas encore de journaux spécialisés dans les *data papers* en linguistique, alors qu'on en trouve dans d'autres domaines, comme l'astronomie, la biologie, la chimie, l'informatique (cf. Dedieu, 2022). Éventuellement, on peut utiliser une revue multidisciplinaire, par exemple le *Journal of Open Humanities Data*[39], dans lequel on peut trouver des articles comme Solopova (2021) dont les données sont déposées sur le site de l'OSF. On peut surtout espérer que la linguistique pourra prochainement se doter de telles revues.

Une ouverture scientifique

44. La mise en place de revues ou de lieux de publication de *data papers* en linguistique est d'autant plus importante qu'elle offre des ouvertures à la linguistique qui peuvent faire très largement évoluer la discipline. En effet, la linguistique est une discipline dans laquelle reproduire des recherches n'est pas aisé, car il y a une grande part d'interprétation dans le travail d'analyse de données langagières. Cette difficulté de reproduction est peut-être à l'origine des grands débats dans la discipline, même si les plus grandes oppositions théoriques reflètent aussi des différences philosophiques de conception du langage et de l'être humain.

45. La citation des données est une des techniques à utiliser pour améliorer la qualité des recherches scientifiques. Elle est aussi un des moyens d'améliorer la reproductibilité des recherches, qui est une des difficultés aujourd'hui rencontrées dans la recherche scientifique (cf. Open Science Collaboration, 2015).

46. Ces avancées devraient se généraliser, car le développement de la science ouverte incite à citer ses données de manière de plus en plus précise. De plus, il semble que la publication de *data papers* et des données qui lui sont associées soit plus efficace pour diffuser ses recherches que la simple publication de corpus (cf. Walters, 2020). Enfin, les données existantes peuvent être utilisées pour d'autres usages que la recherche, comme pour l'enseignement (voir par exemple André, 2019 et le corpus Fleuron[40]). C'est pourquoi il faut espérer que le développement de revues consacrées aux *data papers* pour la linguistique ou ce qui touche au langage sera bientôt une réalité.

38. Voir par exemple https://doranum.fr/plan-gestion-donnees-dmp/fiche-synthetique/ ou https://opidor.fr/category/dmp-faq/.

39. https://openhumanitiesdata.metajnl.com/

40. https://fleuron.atilf.fr/

Conclusion

47. En raison de contraintes naturelles liées à leur coût et à leur intérêt scientifique, les corpus de données de langage, en particulier les corpus de langage oral, et les articles scientifiques qui utilisent ces données, forment un ensemble qui comprend de nombreuses caractéristiques communes avec celles des recommandations que l'on trouve aujourd'hui dans les usages de la science ouverte, de *data papers*, de métadonnées et de données FAIR.

48. Ainsi, les données de langage sont des données brutes, mais les contraintes liées à leurs coûts et leur réutilisation ont fait qu'elles suivent toutes des standards connus et libres de droits. De plus, le dépôt de corpus ne s'accompagne pas de manière obligatoire d'une publication dans une revue scientifique, mais les besoins de la vie d'un chercheur amènent le plus souvent à ce travail de publication. La réutilisation de corpus n'est pas conditionnée à la citation des données, mais le travail des relecteurs scientifiques permet de contrôler cette citation et de demander qu'elle soit respectée. Si elle ne l'est pas, de fait, un article scientifique basé sur des données a peu de valeur. Les sites institutionnels de dépôts de données imposent également l'indication de licence pour la réutilisation des données.

49. Toutes ces contraintes sont apparues de manière relativement naturelle dans le travail des spécialistes de sciences du langage. Le travail d'un consortium comme celui de CORLI vise à documenter ces bons usages et à encourager à leur utilisation. Passer à un usage généralisé de *data papers* associés à des données permettrait très certainement d'améliorer encore les pratiques existantes, de les systématiser et de généraliser les bons usages décrits ci-dessus. L'existence même d'un consortium comme CORLI est la démonstration que la communauté et les instances scientifiques soutiennent cette évolution. Celle-ci serait très certainement bénéfique aux disciplines des sciences du langage, comme elle peut l'être dans d'autres disciplines scientifiques.

Outils et infrastructures

Les *data papers* dans l'écosystème Recherche Data Gouv

Joachim Schöpfel et Christine Kosmopoulos

La mise en place de Recherche Data Gouv

1. Afin de promouvoir une science cumulative et reproductible, les deux plans nationaux de la science ouverte en France ainsi que le programme-cadre Horizon Europe demandent la mise à disposition en accès ouvert des données scientifiques produites dans le cadre des projets de recherche lorsque cela est juridiquement possible. Plus précisément, à côté de la publication des articles en accès ouvert, il s'agit de rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics. Un éventail d'outils, de principes et de protocoles sont mis en place pour permettre à la communauté scientifique de structurer, partager et ouvrir les données de la recherche.

2. L'option d'un entrepôt national de données n'est ni uniforme ni évidente ; il s'agit d'une construction des communautés scientifiques nécessitant une véritable volonté politique (cf. Richou et Schöpfel, 2023). C'est dans ce contexte et après deux études de faisabilité et une étude comparative qu'a été lancé à l'été 2022 l'entrepôt multidisciplinaire Recherche Data Gouv (France) comme « un écosystème au service du partage et de l'ouverture des données »[1].

3. Financé par le Fonds national de la science ouverte (FNSO) à hauteur de 7 610 000 euros, l'écosystème Recherche Data Gouv se donne pour objectif d'offrir la possibilité à tous les chercheurs de pouvoir déposer leurs données sur une plateforme publique, et non commerciale, dès lors que leur communauté ou institution n'en dispose pas. L'idée initiale était de disposer d'un service générique d'accueil et de diffusion des données simples, d'un entrepôt national, multidisciplinaire, mutualisé et public pour les « trous dans la raquette » du paysage scientifique, à l'instar des plateformes Zenodo, Figshare ou DRYAD.

4. Recherche Data Gouv moissonnera également les métadonnées de données partagées via d'autres entrepôts afin de proposer un accès unique aux données de la recherche publique et de faciliter le suivi de la production française dans ce domaine. Un éventail de services accompagne les chercheurs à la transition, pour les aider à transformer leurs pratiques de gestion de leurs données.

1. https://recherche.data.gouv.fr/fr

5. Ainsi, bien plus qu'un simple entrepôt de données, Recherche Data Gouv constitue donc un véritable écosystème public qui se construit autour de cinq éléments (« modules ») :
   - Trois modules pour accompagner les équipes de recherche sur toute question relative aux données :
     - Des ateliers de la donnée ;
     - Des centres de référence thématiques ;
     - Des centres de ressources.
   - Deux modules pour déposer, publier et signaler des données :
     - Un entrepôt pour déposer et utiliser des données ;
     - Un catalogue pour rechercher les données publiées sur cet entrepôt ou sur des entrepôts externes.

6. Au-delà, l'ambition de Recherche Data Gouv est de s'inscrire dans le paysage mondial de la science ouverte en devenant un service de l'European Open Science Cloud (EOSC[2]), offrant un accès au patrimoine des données partagées et ouvertes de la recherche pour favoriser leur réutilisation. Pour mémoire, EOSC est le projet phare du programme-cadre Horizon Europe (2021-2027) visant à proposer un environnement multidisciplinaire fédéré et ouvert où les chercheurs européens pourront publier, mais aussi où les innovateurs, les entreprises et les citoyens trouveront des données ouvertes (y compris des publications), des outils et des services pour la recherche, l'innovation et l'éducation.

7. L'intégration de Recherche Data Gouv dans EOSC est prévue à partir de 2024, avec un financement européen[3]. Parallèlement, l'entrepôt sera connecté à HAL, à deux niveaux : dans le cadre du projet HALiance, HAL développera un service pour le repérage et l'association automatique de la publication déposée dans HAL et les données associées ; et à partir de 2025, il sera possible de déposer des données via un service HAL, en même temps que la publication associée, avec un transfert des données vers Recherche Data Gouv.

8. Au moment de la rédaction de ce chapitre (mai 2023[4]), l'entrepôt contient 544 collections (les « *Dataverses* ») et 2 566 jeux de données, avec au total 31 893 fichiers. De ces jeux de données, 1 940 (76 %) ont été déposés directement sur la plateforme, tandis que les autres (626) ont été moissonnés ailleurs, notamment dans les entrepôts de Sciences Po Paris, de l'IRD et du CIRAD.

9. Côté discipline, trois domaines dominent : l'agriculture (37 %), les sciences de la terre et de l'environnement (27 %) et les sciences de la vie, médecine et santé (19 %). Avec 87 jeux de données, les sciences sociales représentent 3,3 % ; tandis que les arts et les sciences humaines ne représentent que 0,15 % (4 jeux de données).

2. https://eosc-portal.eu/

3. https://recherche.data.gouv.fr/fr/page/trajectoire-recherche-data-gouv

4. https://entrepot.recherche.data.gouv.fr/stats/

10. Au cours des neuf premiers mois de son lancement, l'entrepôt a enregistré 45 460 visiteurs uniques[5]. Au moment de la rédaction, le nombre de téléchargements de fichiers s'élève à 573 630, ce qui correspond à une moyenne arithmétique de 18 téléchargements par fichier de données – un indicateur en hausse depuis 2022.

11. Quant à l'écosystème, le bilan d'avril 2023 fait état de la mise en place de 18 espaces institutionnels sur le portail de l'entrepôt (dont les universités Paris Saclay, Sorbonne Université, Grenoble Alpes et Strasbourg, les organismes INRIA et INRAE et l'Institut Pasteur) et du lancement de 15 ateliers de la donnée.

12. Par la suite, nous allons voir la place des *data papers* dans cet écosystème Recherche Data Gouv.

Les *data papers* dans les ateliers de la donnée

13. Les ateliers de la donnée sont mis en place progressivement sur le territoire, avec une labellisation et un soutien financier du Fonds national pour la science ouverte (FNSO). Leur mission est d'être un « point d'entrée facilement identifiable par les équipes de recherche pour toute question relative à la gestion des données de la recherche », avec des « services généralistes d'appui aux problématiques de gestion, structuration, partage et ouverture des données de recherche ». Il s'agit de dispositifs de « mise en commun de ressources et compétences pour apporter à proximité des chercheurs, un premier niveau d'expertise, sur toute problématique de l'ensemble du cycle de vie de la donnée », en complémentarité des services d'accompagnement déjà existants[6].

14. Le cahier des charges de ces ateliers évoque les *data papers* comme l'une de leurs missions, en ces termes précis : « Accompagner la rédaction de *data papers*/articles de données : conseils de rédaction, choix de la revue appropriée… » (*idem*). Les établissements qui veulent mettre en place un tel atelier sont donc invités à se positionner par rapport aux *data papers* et *data journals*, en s'appuyant sur une offre de service existante (par exemple celle mise en place par les Urfist [*Unité régionale de formation à l'information scientifique et technique*]), et/ou en proposant de nouvelles prestations de conseil, d'information et/ou de formation.

15. La réalité sur le terrain est contrastée et évolutive. Plusieurs ateliers ont déjà intégré les *data papers* dans leur offre de service, d'autres en font l'annonce[7]. Cette prestation prend le plus souvent la forme de l'accompagnement, c'est-à-dire de la facilitation (aider à faire) et

5. Bilan des neuf premiers mois https://recherche.data.gouv.fr/fr/actualite/recherche-data-gouv-a-9-mois.

6. Cf. Appel à manifestation d'intérêt n° 1 « Ateliers de la donnée » 3 février 2022 https://recherche.data.gouv.fr/fr/page/appels-a-manifestation-dinteret-ateliers-de-la-donnee.

7. Ateliers de la donnée https://recherche.data.gouv.fr/fr/page/ateliers-de-la-donnee-des-services-generalistes-sur-tout-le-territoire.

de la médiation (apprendre à faire) plus que de la coproduction (faire avec).

16. Ainsi, les ateliers des universités Paris Saclay, Rennes et Lorraine affichent explicitement l'accompagnement à la rédaction et la publication d'un *data paper* comme l'une de leurs missions, sans préciser les modalités de cet accompagnement. La Cellule Data Grenoble Alpes propose la relecture d'un *data paper*, tandis que l'atelier de l'Université Gustave Eiffel est prêt à conseiller sur le choix de revues pour la publication d'un *data paper*.

17. L'Université de Reims Champagne-Ardenne a mis en ligne une page sur la rédaction d'un *data paper*, similaire à l'Université de Bourgogne Franche-Comté où l'on trouve une page sur la valorisation des données, avec des liens vers quatre supports d'information et de formation, des ressources mises en ligne sur les sites de DoRANum du CNRS[8] et de CoopIST du CIRAD[9].

18. Il faudra voir comment ces ateliers vont s'articuler avec l'offre d'accompagnement des centres de référence établissements déjà en place, comme l'INED ou Sciences Po[10].

La place des *data papers* dans les centres de référence thématiques

19. Plusieurs infrastructures, labellisées « centres de référence thématiques », soutiennent la gestion, le traitement, le partage et l'ouverture des données dans un domaine scientifique particulier. Elles sont impliquées dans la standardisation et les bonnes pratiques de leurs communautés de données, font le lien avec l'international et soutiennent l'articulation entre les dispositifs thématiques spécialisés et Recherche Data Gouv.

20. Début 2023, l'écosystème compte six centres de référence thématiques, dans quatre grands domaines scientifiques (astronomie et astrophysique, sciences de la terre et de l'environnement, SHS, biologie et santé)[11]. Chaque centre a sa particularité, avec un positionnement et une offre de service propres à sa communauté. Aussi, la place des *data papers* dans ces centres n'est pas systématique :

- L'entrepôt de données de l'infrastructure de recherche Data Terra annonce la génération de *data papers* « dans un second temps »[12].
- Le Pôle National de Données de Biodiversité (PNDB) mène actuellement le projet OpenMetaPaper, coordonné par le Muséum national d'histoire naturelle et financé par le FNSO. En lien avec le GBIF (cf. le chapitre de Pamerlon dans ce livre), OpenMetaPaper a comme objectif de « booster l'ouverture des données

8. https://doranum.fr/
9. https://coop-ist.cirad.fr/
10. Centres de référence établissements https://recherche.data.gouv.fr/fr/page/centres-de-reference-etablissements.
11. Centres de référence thématiques https://recherche.data.gouv.fr/fr/page/centres-de-reference-thematiques-expertises-par-domaine-scientifique.
12. https://www.data-terra.org/donnees-services/entrepot-de-donnees-data-terra/

de recherche en écologie », mettant l'accent « sur la publication scientifique, objet de recherche principal de la valorisation des activités scientifiques, en testant un dispositif permettant de (1) booster la production de "*data paper*" par la communauté en écologie et (2) augmenter l'impact de ces articles en facilitant la publication de tel matériel dans des revues à haut facteur »[13].

- En SHS, l'entrepôt Nakala de Huma-Num contient quelques *data papers* et des liens vers des *data papers*[14].

21. Par ailleurs, et pour la suite, il paraît logique que le travail de normalisation de ces centres de référence thématiques porte également sur l'interconnexion des standards et formats des entrepôts de données et des plateformes de revues, avec un impact sur la production et la publication des *data papers*.

Les *data papers* dans les centres de ressources

22. De différentes natures, les centres de ressources nationaux apportent un appui aux ateliers de la donnée et aux centres de référence thématiques. Recherche Data Gouv fédère des ressources et des services existants pour permettre une capitalisation à l'échelle nationale et au bénéfice de tous[15]. Actuellement, l'écosystème regroupe quatre centres de ressources :

23. Entrepôt-catalogue : Ce centre de ressources est directement lié à l'entrepôt. Il assure la création d'espaces institutionnels, l'assistance et la formation des administrateurs et curateurs des espaces institutionnels, des ateliers de la donnée et des équipes de recherche dont les établissements ne disposent pas d'espaces institutionnels, et la modération et la curation des jeux de donnée n'étant pas dans un espace institutionnel.

24. Compétences : Porté par le GIS « Réseau URFIST », ce deuxième centre de ressources organise et développe des modules de formations dédiés aux ateliers de la donnée. Parmi les formations accessibles sur la plateforme Callisto, on trouve les *data papers* à deux endroits : comme un moyen pour trouver des données et (surtout) comme une voie formalisée pour la publication et le partage des données, une manière de rendre les données accessibles, interprétables et réutilisables. Ces formations s'adressent aux doctorants et aux chercheurs, mais aussi aux professionnels de l'information et aux formateurs.

25. DoRANum : Le portail DoRANum de l'Inist-CNRS et du GIS « Réseau URFIST » propose un large éventail de formations à distance sur la thématique de la gestion et du partage des données de recherche, des guides, modules de formation, fiches synthétiques, conférences, infographies, tutoriels, etc., au total 275 ressources (mai 2023) qui servent de support à l'offre de service des établissements, organismes, ateliers de la donnée, etc. Comme déjà évoqué plus haut, les *data papers* font bien entendu partie de cette offre de formation à distance du

13. https://www.pndb.fr/fr/projets/opemetadatapaper-un-projet-fond-national-pour-la-science-ouverte
14. https://nakala.fr/
15. Centres de ressources https://recherche.data.gouv.fr/fr/page/centres-de-ressources.

portail DoRANum, notamment sous forme d'une fiche synthétique sur les *data papers* et les *data journals* (DOI : 10.13143/2WCB-FW52), d'une présentation du contenu d'un *data paper* (DOI : 10.13143/8R3D-K505), d'une autre fiche synthétique sur les critères d'évaluation d'un *data paper* (DOI : 10.13143/DK71-W 757) ou encore d'un webinaire du GTSO Données du consortium Couperin avec trois retours d'expériences de publication d'un *data paper* (DOI : 10.13143/SX95-CF48). Plusieurs de ces ressources font l'objet d'une diffusion sur Canal U dont en particulier une conférence de Laurence Dedieu du Cirad sur les enjeux de la publication d'un *data paper*[16].

26. OPIDoR : Le quatrième centre de ressources, le portail OPIDoR, a été mis en place par l'Inist-CNRS afin de mettre à disposition de la communauté de l'Enseignement Supérieur et de la Recherche un ensemble d'outils et de services pour une bonne gestion des données de recherche (FAIRisation). Son offre consiste en trois services, et les trois services ont un lien avec les *data papers* :

- DMP-OPIDoR : L'outil d'aide à la création en ligne de plans de gestion de données facilite la rédaction d'un *data paper* par la description riche et structurée des jeux de données, avec des rubriques qui correspondent en grande partie aux contenus des *data papers*. En plus, dû à son caractère « machine actionable », DMP-OPIDoR peut être connecté à des outils de rédaction.
- Cat-OPIDoR : Sous forme d'un wiki, ce catalogue recense et décrit les services français qui contribuent à la gestion, au partage et la réutilisation des données de recherche. Il ne contient pas d'inventaire de *data papers*, mais, parmi les services recensés, se trouvent plusieurs qui eux proposent une offre dédiée aux *data papers*, comme les ateliers de la donnée des universités Paris Saclay et Lorraine (cf. *infra*) ou l'entrepôt Cybergeo Dataverse. Cependant, cette description n'est pas exhaustive ; il manque par exemple l'accès aux *data papers* sur le portail d'accès aux données sur la biodiversité produit par le GBIF France[17], avec en plus la possibilité de générer des *data papers* à partir des métadonnées des données (cf. Pamerlon dans ce livre).
- PID-OPIDoR : Conçu initialement comme un service d'attribution d'un identifiant pérenne (le DOI de DataCite) aux jeux de données, PID-OPIDoR permet aussi, surtout après l'implémentation de la version 4.4 du schéma de métadonnées de DataCite, l'attribution d'un DOI à d'autres résultats de recherche, y compris les *data papers*[18].

27. La description de ces différents centres de ressources révèle leur caractère d'écosystème, l'interconnexion de leur offre de service. Dans ce système, les *data papers* n'occupent peut-être pas une place prépondérante ; néanmoins ils sont clairement identifiés comme un objet d'information, de formation, d'assistance et de conseil, avec également, mais plus en marge, une dimension technologique (attribution d'un identifiant, réutilisation des plans de gestion, etc.).

16. Publier un data paper https://www.canal-u.tv/chaines/callisto/publier-un-data-paper-laurence-dedieu.

17. https://cat.opidor.fr/index.php/GBIF_portail_France

18. Cf. l'exemple https://support.datacite.org/docs/schema-metadata-examples-v41#example-for-a-data-paper.

La place des *data papers* dans l'entrepôt

28. Cette dimension technologique est en revanche tout à fait visible et présente dans le cœur de l'écosystème, c'est-à-dire, dans l'entrepôt Recherche Data Gouv lui-même. En effet, l'entrepôt national a intégré une application JAVA développée par l'INRAE pour son propre entrepôt de données : la génération automatique d'un *data paper* à partir des métadonnées d'un jeu de données, identifié par son DOI.

29. Cette fonctionnalité est accessible sur la page d'accueil de l'entrepôt[19] ; elle permet de créer « une ébauche de *data paper* » à partir du DOI d'un jeu de données déposé dans l'entrepôt[20]. Actuellement, cette génération est possible dans deux formats :

- Recherche Data Gouv : extraction de toutes les métadonnées du jeu de données, en format texte (odt), avec une information bibliographique structurée (header), dont le titre, les auteurs avec leur affiliation et identifiant (ORCID), un résumé avec des mots-clés, le domaine scientifique et le type de données (figure 1). Quant à la qualité des autres informations, l'ébauche de ce *data paper* sera aussi riche que l'indexation du jeu de données par le(s) contributeur(s). Autrement dit, l'entrepôt ne dispose pas (encore ?) d'un service d'indexation automatique pour enrichir les métadonnées créées (ou importées) au moment du dépôt (ou de l'import) des données.
- *Data in Brief* : extraction des métadonnées demandées par la revue *Data in Brief*[21] d'Elsevier, mise en forme selon les instructions aux auteurs, en format texte (odt). La figure 2 montre le *data paper* du même jeu de données, mais avec le modèle (*template*) de *Data in Brief*. Les informations bibliographiques sont suivies par le résumé ; quant aux autres rubriques, notamment le tableau des spécifications du *template*, elles restent à remplir ou à compléter (figure 3).

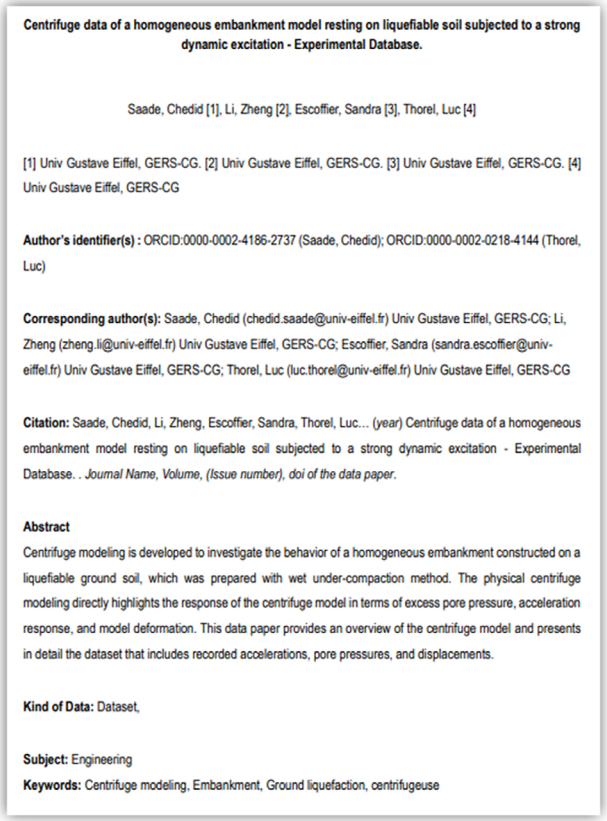
30. Ce service fonctionne pour tous les jeux de données déposés dans Recherche Data Gouv, y compris pour ceux qui ne sont pas publiés (accès limité). Comme déjà évoqué, la qualité du résultat dépend essentiellement de la qualité et de la richesse des métadonnées. Lors d'une réunion de la *Dataverse Community* (4 avril 2023), d'autres limitations ont été évoquées : l'absence de conditions d'utilisation pour les métadonnées (pas de licence), une visibilité limitée (car le service n'est pas accessible au niveau des jeux de données, seulement sur la page d'accueil de l'entrepôt), une procédure non intuitive (saisie du DOI dans un champ, à la place d'un simple bouton de service au niveau des données), des rubriques et instructions en anglais, et seulement deux formats. Une traduction serait en préparation ; quant aux formats, d'autres *templates* pourraient être intégrés, permettant ainsi la création d'ébauches pour d'autres revues de données.

31. Contrairement à d'autres entrepôts, comme Zenodo ou Figshare, Recherche Data Gouv ne prévoit pas le dépôt

19. https://entrepot.recherche.data.gouv.fr/dataverse/root

20. Création d'un data paper https://entrepot.recherche.data.gouv.fr/datapartage-datapapers-web/.

21. https://www.sciencedirect.com/journal/data-in-brief



Centrifuge data of a homogeneous embankment model resting on liquefiable soil subjected to a strong dynamic excitation - Experimental Database.

Saade, Chedid [1], Li, Zheng [2], Escoffier, Sandra [3], Thorel, Luc [4]

[1] Univ Gustave Eiffel, GERS-CG. [2] Univ Gustave Eiffel, GERS-CG. [3] Univ Gustave Eiffel, GERS-CG. [4] Univ Gustave Eiffel, GERS-CG

Author's identifier(s) : ORCID:0000-0002-4186-2737 (Saade, Chedid); ORCID:0000-0002-0218-4144 (Thorel, Luc)

Corresponding author(s): Saade, Chedid (chedid.saade@univ-eiffel.fr) Univ Gustave Eiffel, GERS-CG; Li, Zheng (zheng.li@univ-eiffel.fr) Univ Gustave Eiffel, GERS-CG; Escoffier, Sandra (sandra.escoffier@univ-eiffel.fr) Univ Gustave Eiffel, GERS-CG; Thorel, Luc (luc.thorel@univ-eiffel.fr) Univ Gustave Eiffel, GERS-CG

Citation: Saade, Chedid, Li, Zheng, Escoffier, Sandra, Thorel, Luc… (*year*) Centrifuge data of a homogeneous embankment model resting on liquefiable soil subjected to a strong dynamic excitation - Experimental Database. . *Journal Name, Volume, (Issue number), doi of the data paper.*

Abstract

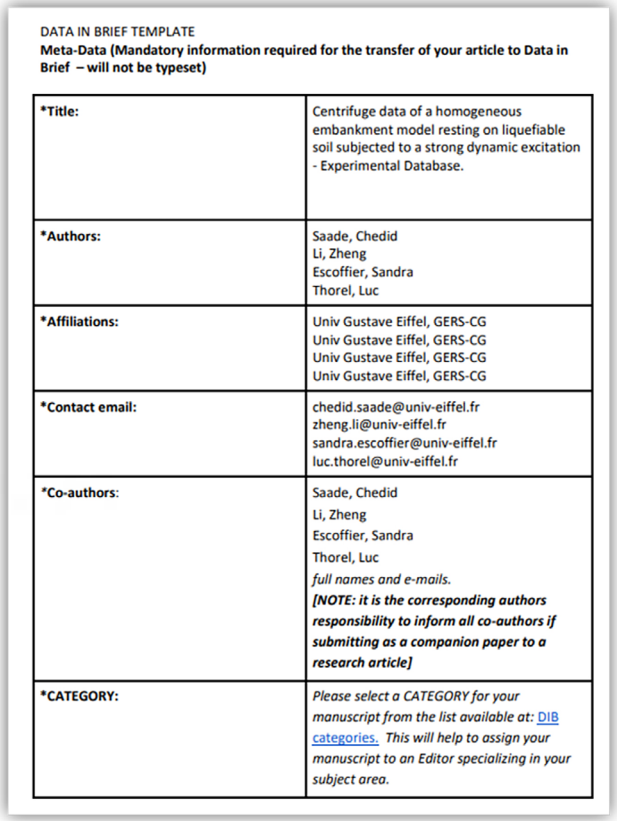
Centrifuge modeling is developed to investigate the behavior of a homogeneous embankment constructed on a liquefiable ground soil, which was prepared with wet under-compaction method. The physical centrifuge modeling directly highlights the response of the centrifuge model in terms of excess pore pressure, acceleration response, and model deformation. This data paper provides an overview of the centrifuge model and presents in detail the dataset that includes recorded accelerations, pore pressures, and displacements.

Kind of Data: Dataset,

Subject: Engineering

Keywords: Centrifuge modeling, Embankment, Ground liquefaction, centrifugeuse

Figure 1. Exemple d'un début de *data paper* généré par Recherche Data Gouv (DOI : 10.57745/LDNZR9 ; format Recherche Data Gouv)

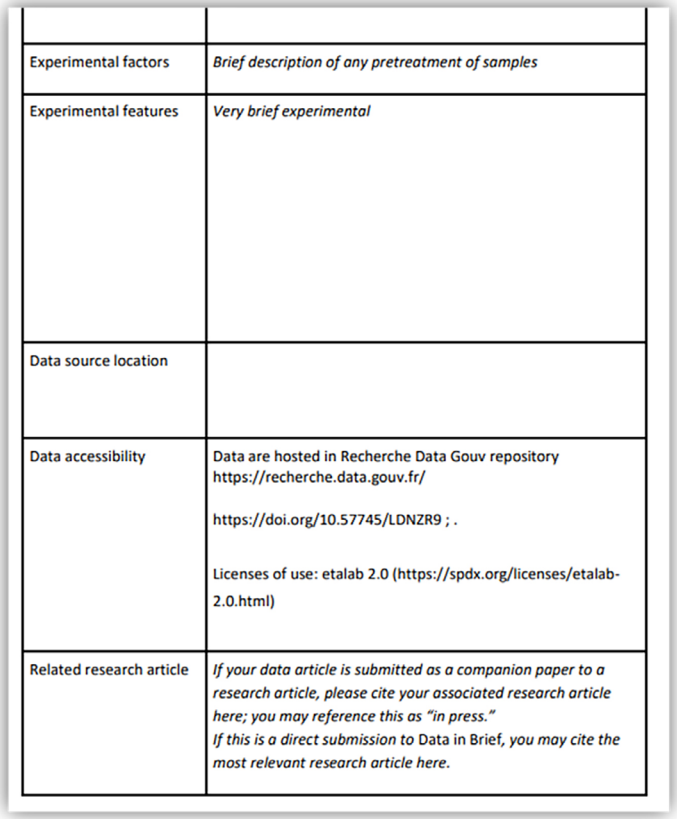


DATA IN BRIEF TEMPLATE

Meta-Data (Mandatory information required for the transfer of your article to Data in Brief – will not be typeset)

***Title:**	Centrifuge data of a homogeneous embankment model resting on liquefiable soil subjected to a strong dynamic excitation - Experimental Database.
***Authors:**	Saade, Chedid Li, Zheng Escoffier, Sandra Thorel, Luc
***Affiliations:**	Univ Gustave Eiffel, GERS-CG Univ Gustave Eiffel, GERS-CG Univ Gustave Eiffel, GERS-CG Univ Gustave Eiffel, GERS-CG
***Contact email:**	chedid.saade@univ-eiffel.fr zheng.li@univ-eiffel.fr sandra.escoffier@univ-eiffel.fr luc.thorel@univ-eiffel.fr
***Co-authors**:	Saade, Chedid Li, Zheng Escoffier, Sandra Thorel, Luc *full names and e-mails.* ***[NOTE: it is the corresponding authors responsibility to inform all co-authors if submitting as a companion paper to a research article]***
***CATEGORY:**	*Please select a CATEGORY for your manuscript from the list available at:* DIB categories. *This will help to assign your manuscript to an Editor specializing in your subject area.*

Figure 2. Exemple d'un début de *data paper* généré par Recherche Data Gouv (DOI : 10.57745/LDNZR9 ; format *Data in Brief*)



Experimental factors	*Brief description of any pretreatment of samples*
Experimental features	*Very brief experimental*
Data source location	
Data accessibility	Data are hosted in Recherche Data Gouv repository https://recherche.data.gouv.fr/ https://doi.org/10.57745/LDNZR9 ; . Licenses of use: etalab 2.0 (https://spdx.org/licenses/etalab-2.0.html)
Related research article	*If your data article is submitted as a companion paper to a research article, please cite your associated research article here; you may reference this as "in press."* *If this is a direct submission to* Data in Brief*, you may cite the most relevant research article here.*

Figure 3. Extrait du *specification table* du même *data paper* – en italique, les instructions de la revue (DOI : 10.57745/LDNZR9 ; format *Data in Brief*)

et le partage des *data papers* sur la plateforme, avec les données de recherche. Comme indiqué plus haut, l'association entre données et *data papers* se fera probablement via les services à développer sur HAL et par le biais d'identifiants pérennes.

32. Néanmoins, il paraît techniquement tout à fait possible de déposer le fichier d'un *data paper* avec le jeu de données – en fait, comparable à un fichier ReadMe en format texte, comme sur l'entrepôt néerlandais EASY de l'institut Data Archiving and Networked Services (DANS)[22].

La connexion avec les revues – l'exemple de *Cybergeo*

33. En 2017, la revue *Cybergeo : revue européenne de géographie*[23] lançait une rubrique dédiée aux *data papers*. La soumission d'un *data paper* suit la même procédure que les autres articles : les auteurs sont invités à soumettre de manière anonyme leur article sur le site de soumission en ligne de la revue. L'application utilisée pour les soumissions de *Cybergeo* est OJS, une application en open source développée par le Public Knowledge Project (PKP) de l'Université de Vancouver[24]. Cette application avait été proposée un temps par OpenEdition au début des années 2010 pour la gestion du workflow, puis la plateforme française a cessé de proposer ce type de service. *Cybergeo* a alors entamé un partenariat directement avec le PKP qui héberge le site de soumissions et permet au comité de rédaction de gérer l'évaluation des différentes versions en double aveugle.

34. Dans la version la plus récente du logiciel (OJS 3.3.0.11), il est possible d'associer à l'application un certain nombre de plug-ins développés par la communauté OJS, mais à ce stade, il n'y a pas moyen d'associer les données à un *data paper* lors de la soumission. La politique éditoriale adoptée est donc de soumettre d'un côté le *data paper* sur la plateforme OJS et de l'autre de déposer les données dans un entrepôt institutionnel de données qui réponde aux principes FAIR, de préférence dans l'un des trois entrepôts où se trouvent les collections de la revue. En effet, la revue *Cybergeo* réunit les jeux de données publiés en association avec les *data papers* dans des collections dans l'un des trois entrepôts : Nakala, Zenodo et Cybergeo Dataverse[25]. Le lien vers l'entrepôt et le DOI de publication sont insérés dans l'article principal publié dans la revue. L'un des contributeurs de la communauté de l'application en open source *Dataverse*, utilisée par Recherche Data Gouv, travaille actuellement sur un plug-in de connecteur Dataverse[26] qui permettra aux auteurs de créer un lien vers l'ensemble de leurs données lors du processus de soumission. Un plug-in pour OJS devrait être proposé prochainement

22. Un exemple : https://doi.org/10.17026/dans-x7y-wmyc.
23. Cybergeo est hébergée sur la plateforme OpenEdition https://journals.openedition.org/cybergeo/.
24. https://pkp.sfu.ca/software/ojs/
25. Cybergeo Dataverse sur la plateforme d'Harvard https://dataverse.harvard.edu/dataverse/cybergeo.
26. Dataverse plugin https://github.com/lepidus/dataversePlugin.

par les développeurs. Cette fonctionnalité facilitera grandement le traitement par les équipes éditoriales des soumissions des *data papers* et de leur publication, pas seulement pour la revue *Cybergeo*, mais aussi pour d'autres revues qui publieraient des *data papers* avec des données déposées sur Recherche Data Gouv.

L'enjeu de l'évaluation en double aveugle d'un *data paper*

35. Un autre problème rencontré par les équipes éditoriales dans la gestion du *data paper* est le respect de l'évaluation en double aveugle. En effet, la spécificité du contenu des *data papers* impose non seulement des contraintes en termes de publication, mais aussi en termes d'évaluation que ne rencontrent pas les articles classiques. Plus précisément, les *data papers* en tant qu'articles de données sont soumis à des revues et à l'évaluation en double aveugle par les pairs tout comme l'article scientifique classique, mais l'association aux données déposées dans un entrepôt pose un problème supplémentaire. Les pratiques dans la procédure d'évaluation sont très variables d'une discipline à l'autre, d'une revue à l'autre, parce qu'il n'existe pas de protocoles partagés ni d'entrepôt adapté à l'anonymisation des jeux de données des *data papers* lors de la soumission. Or, la double anonymisation de l'examen par les pairs des contenus des articles et du matériel supplémentaire (codes, données, etc.) est une exigence si l'on tient à respecter les critères scientifiques et éthiques qu'on s'efforce d'appliquer depuis des décennies aux articles scientifiques classiques.

36. À travers l'évaluation par les pairs, le *data paper* est le garant de la validation scientifique des jeux de données directement publiés dans les entrepôts de données. Or, force est de constater qu'aucun entrepôt n'est à ce jour adapté aux procédures d'évaluation pour les *data papers*. L'éditeur se trouve alors contraint de détourner certaines fonctionnalités des entrepôts institutionnels, comme dans le Dataverse utilisé par Recherche Data Gouv, pour respecter l'anonymisation des auteurs lors de la soumission de leurs données associées à leur article et des évaluateurs lorsqu'ils accèdent à ces données pour rendre un avis sur le *data paper*.

37. Nous prenons ici à titre d'exemple à nouveau les difficultés rencontrées par la revue *Cybergeo : revue européenne de géographie*. La revue a lancé en 2014 une rubrique *Model papers* et en 2017 une rubrique *Data papers*. L'objectif de ces articles est de décrire, documenter et partager les bases de données et les codes sources des modèles reproductibles produits par les auteurs afin de permettre leur réutilisation et d'avancer vers une science reproductible. Chacune de ces deux rubriques est dotée d'un comité de lecture propre d'une vingtaine d'experts du domaine, qui ont pour obligation d'évaluer en double aveugle les articles et d'avoir accès au matériel supplémentaire qui les accompagne (voir chapitre de Cottineau-Mugadza *et al.* dans ce livre).

38. Le matériel supplémentaire comporte des jeux de données associés aux *data papers* qui sont déposés dans un entrepôt de données selon la procédure indiquée dans la partie précédente. Ce matériel répond aux mêmes exigences que l'article principal dans toutes les étapes de la procédure d'évaluation depuis la soumission jusqu'à l'acceptation, c'est-à-dire que les évaluateurs ne doivent pas avoir connaissance des auteurs de l'article. Lors de la soumission, il est donc demandé aux auteurs de déposer anonymement leur article sur le site de soumission de la revue ainsi que leurs données dans un entrepôt institutionnel pérenne où les données pourront être publiées après acceptation. En effet, pour respecter les règles de l'évaluation par les pairs, les évaluateurs doivent pouvoir contrôler les données, faire fonctionner les modèles proposés, vérifier la reproductibilité des expériences qui ont été effectuées, et cela sans connaître l'identité de leurs auteurs. On est alors surpris de découvrir que, bien que des systèmes aient été conçus pour protéger l'anonymat des données personnelles, la possibilité d'appliquer cette clause aux auteurs et aux évaluateurs des *data papers* n'a, elle, pas été anticipée alors qu'il s'agit d'un principe *sine qua none*.

39. *Cybergeo* dispose de plusieurs collections de données sur des entrepôts institutionnels de données nationaux et internationaux (Nakala, Zenodo, Dataverse), mais aucun de ces entrepôts institutionnels ne dispose de la fonctionnalité d'anonymisation des données. Plus encore, lors du dépôt des données, certains entrepôts génèrent automatiquement un DOI public qui permet d'accéder aux données et à toutes les informations sur les auteurs en naviguant sur le Net, sans toutefois que ces données aient été validées par les évaluateurs du *data paper*.

40. Dans Dataverse, il est toutefois possible de générer un DOI provisoire sans publication, et de le supprimer si le *data paper* correspondant n'est pas accepté. Cependant, le modèle qui encadre le dépôt des métadonnées ne permet pas l'anonymisation et donc, inévitablement l'évaluateur aura connaissance du nom des auteurs, si ce n'est par l'article soumis, par les données déposées dans l'entrepôt. Dans ce cas précis, il est demandé aux auteurs de remplir les champs « Nom » par le numéro de soumission de l'article et de supprimer toutes les données qui pourraient les identifier. Lorsque l'article est accepté, toutes les métadonnées doivent être mises à jour avant publication.

41. Cette absence de fonctionnalité à l'adresse des revues publiant des *data papers* dans les entrepôts de données pose un vrai problème. On le retrouve également dans le Dataverse de Recherche Data Gouv. Confrontés à cette difficulté, un projet *Cybergeo Science Reproductible* (CSR) avait été soumis en 2021 au FNSO qui avait pour objectif des propositions d'adaptation des entrepôts de données à l'évaluation en double aveugle des *data papers* avec des fonctionnalités répondant aux besoins spécifiques des différents acteurs : auteurs, évaluateurs, éditeurs, rédacteurs, etc., qui interviennent tout au long du processus éditorial et de mise en partage de protocoles de soumission, d'évaluation et de publication des données dans le contexte d'un *data paper*.

42. Simultanément un partenariat a été mis en place avec Harvard Dataverse pour développer une fonctionnalité d'anonymisation des données pour les auteurs et les évaluateurs lors de l'évaluation d'un *data paper*. Les différents échanges entre *Cybergeo* et la plateforme ont conduit à des tests pertinents avec un bouton qui rend invisibles certains champs selon la façon dont on se connecte à la plateforme. Bien qu'efficace, la fonctionnalité n'a toutefois pas encore été mise en place de façon pérenne.

Perspectives

43. On observe bien un nombre croissant de *data papers*, c'est-à-dire d'articles qui décrivent, documentent et mettent en partage des jeux de données. Toutefois, si l'entrepôt de données est indispensable au *data paper*, force est de constater que les entrepôts de données n'ont pas encore adapté leur offre de services aux spécificités de ce type de publication, tant sur le plan de la soumission que de l'évaluation.

44. Néanmoins, l'étude du nouveau dispositif Recherche Data Gouv montre que les *data papers* ont bien leur place dans tous les modules de cet écosystème. Comme ce système est évolutif et modulaire, il est plus que probable que d'autres services voient le jour, pour la génération des *data papers*, pour leur rédaction, leur soumission et leur évaluation, peut-être aussi pour le dépôt et la publication. Cela dépendra sans doute de l'agilité du dispositif, mais surtout du développement des pratiques dans les communautés, établissements et organismes et aussi, de l'acceptation des *data papers* par les équipes scientifiques. Quant à Recherche Data Gouv, vu sa trajectoire, une question reste pour l'instant ouverte : le futur catalogue, qui sera mis en place progressivement pour signaler et moissonner les métadonnées de jeux de données publiées sur des entrepôts externes, contiendra-t-il aussi des liens vers des *data papers* ?

Vers un écosystème d'écriture et d'édition avec les données

Nicolas Sauret, Stéphane Pouyllau et Mélanie Bunel

Constat et contextualisation

1. Depuis 2013, l'infrastructure Huma-Num[1] propose aux communautés de recherche des sciences humaines et sociales (SHS) un ensemble de services et d'outils qui tendent à esquisser un environnement de travail dédié aux équipes de recherche de ces disciplines. L'histoire de cette infrastructure (Pouyllau *et al.*, 2021) rend compte d'une évolution des services au plus près des besoins de la communauté SHS, fortement engagée dans la co-conception des outils de la recherche.

2. La décennie 2010-2020 a vu s'accélérer encore l'appropriation à la fois des enjeux et des outils numériques par les communautés SHS (Clavert et Schafer, 2019). Cependant, si les infrastructures de recherche et le mouvement des humanités numériques se sont très fortement développés[2] – et le mouvement est mondial (« Nouvelles perspectives sur les humanités numériques », 2021) – se pose plus que jamais la question de la littératie numérique au cœur des programmes de recherche. La seconde moitié de la décennie 2010-2020 est marquée par deux phénomènes. D'un côté, les infrastructures de recherche ont évolué vers le rôle de « super DSI[3] », palliant par exemple les difficultés des DSI d'établissements universitaires à absorber les volumes de plus en plus importants de données, ainsi que la nécessité de proposer des services interconnectés à l'échelle mondiale et « sans couture » dans un cloud universitaire mondial. De l'autre côté, le développement des outils numériques au cœur même des programmes de recherche a été facilité grâce à une plus grande accessibilité à des ressources de développement, les développeurs fonctionnant de plus en plus en freelance. Par ailleurs, la démocratisation des plateformes de codage[4] et des framework de développement a déplacé la forge de l'outillage numérique au plus près des programmes de recherche. De plus en plus de chercheurs et de chercheuses font l'expérience du co-développement et/ou du codage d'applications légères de traitement des

1. Huma-Num a reçu, dès 2013, le label de « très grande infrastructure de recherche » (ou TGIR) du ministère de la Recherche, de l'Enseignement supérieur et de l'Innovation.

2. Par exemple, avec la création d'autres infrastructures nationales (Progedo, Metopes, Open Edition, etc.) et européennes (Dariah, Clarin, Operas, Resilence, etc.). Mais aussi avec l'institutionnalisation des humanités numériques (master, doctorat, postes d'enseignants universitaires mis au recrutement, association professionnelle, revues, etc.

3. Direction des systèmes d'information.

4. On pense notamment aux plateformes de *data science*, avec par exemple JupyterLab.

données. La banalisation et l'appropriation de langages tels que R ou Python en ont ainsi accéléré le processus. Cependant, ce mouvement n'est pas totalement massif puisque des communautés restent encore éloignées de ces transformations des pratiques de recherche.

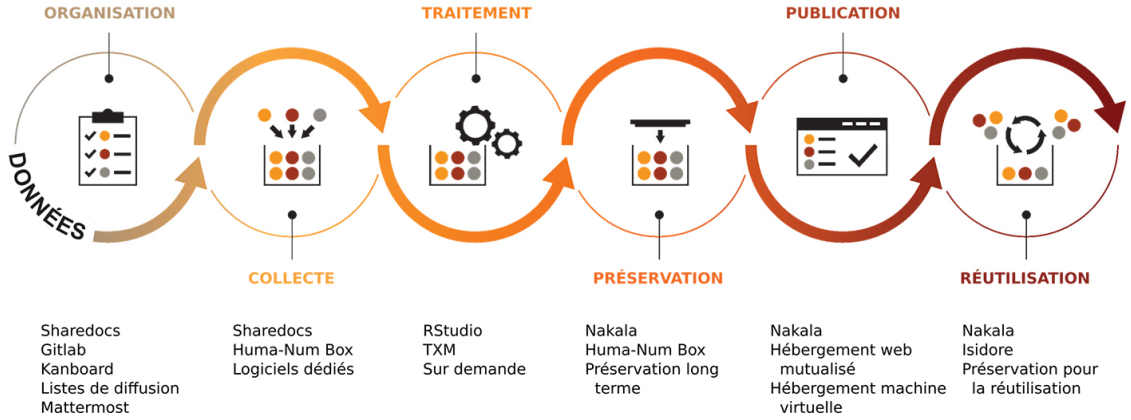


Figure 1. Les outils et services de l'infrastructure Huma-Num dans le cycle de la donnée

Crédit : doi:10.34847/nkl.54afy3ed/d3793bc8ae69513f58cdc8acf6181b1a23fd31ab - Creative Commons Attribution Share Alike 4.0 International (CC-BY-SA-4.0)

3. Ces deux phénomènes créent aussi de nouveaux questionnements : quelle pérennité pour ces programmes et pour l'outillage de la recherche en général ? Comment encourager la réflexivité si essentielle aux SHS sur les programmes, sur les modalités de traitement mis en œuvre sur les données et sur le cœur même des dispositifs socio-techniques créés (Akrich, 1992) ? Comment passer d'une expérimentation de traitement de données à l'échelle individuelle (par exemple sous R ou sur une instance de Jupyter fonctionnant sur son ordinateur individuel) à une échelle plus collective, c'est-à-dire celle des pairs, et plus stable dans le temps, celle de la publication scientifique dans le temps long ? Il nous semble que ces deux enjeux de réflexivité sur les traitements et de publicisation du processus de recherche redéfinissent la place, les services et les organisations des infrastructures de recherche telles qu'elles se sont développées entre 2005 et 2007. C'est dans ce contexte d'évolution permanente et parfois rapide du développement du numérique dans la recherche en SHS, que Huma-Num a pu se doter d'un laboratoire de recherche et d'innovation.

4. Le HN Lab[5], créé en juin 2020 au sein d'Huma-Num, est aujourd'hui engagé dans une réflexion visant à implémenter une vision plus intégrée des infrastructures vers un véritable écosystème de recherche intégrant l'ensemble du cycle des données de recherche et en l'adaptant aux pratiques des disciplines SHS. Cette vision remet les écritures numériques au cœur des activités des chercheurs et chercheuses (Houot et Triby, 2020), comprises comme l'expression et la diffusion des savoirs. Les formats *data papers* et *executable papers* s'inscrivent dans cette vision comme des modèles éditoriaux susceptibles d'accueillir les changements de pratiques évoqués.

5. Voir https://www.huma-num.fr/hnlab.

5. Outre l'intégration technique des différentes plateformes constitutives de cet écosystème, notre réflexion porte sur le modèle épistémologique sous-jacent. Nous nous intéressons à la fois aux modalités de production et de diffusion du savoir, mais aussi aux modalités d'appropriation et d'accès à ce modèle. Il s'agit en fait de concevoir ensemble[6] l'écosystème et la littératie numérique de la donnée et de l'écriture scientifique. De ce point de vue, notre hypothèse est celle d'une appropriation progressive de l'écosystème et de ses plateformes, et donc d'une inclusion des savoirs techniques et méthodologiques sur lesquels ils reposent, favorisée par une convergence et une interrelation des diverses pratiques d'écriture qui composent l'activité scientifique.

Repenser l'écosystème de recherche

6. Notre vision inscrit les données de la recherche au cœur d'un écosystème de recherche tourné vers leur production, leur circulation et leur réutilisation, favorisant l'adoption des principes de la science ouverte, en forte effervescence depuis quelques années[7].

7. Dans l'imaginaire des métiers de la recherche, la donnée est souvent conceptualisée à partir de son cycle de vie. Nous sommes maintenant familiers de ces différentes représentations schématisant le circuit générique de la donnée. Mais les pratiques de recherche nous obligent à déconstruire cette conceptualisation du cycle de la donnée. À l'épreuve du terrain, on comprend en effet assez vite les limites de tels schémas, en particulier leur aspect séquentiel. De fait, l'activité de recherche ne peut pas se réduire à une succession d'étapes. Plus objectivement, il serait en fait plus juste de parler d'états de la recherche plutôt que d'étapes, considérant en effet que les représentations du « cycle de la donnée » établissent trop schématiquement une séquence d'actions là où les pratiques tendent désormais à procéder simultanément aux différentes actions. L'un des fondements guidant cette vision consiste à envisager la donnée comme étant engagée dans une multiplicité d'écritures, qu'elles soient de l'ordre de l'encodage, du code logiciel et algorithmique, de la rédaction ou de l'édition, outillées chacune par des dispositifs dédiés selon les états de la recherche.

8. Cette vision propose de déplacer le regard (zoom-in) et de 1) considérer le processus de recherche comme une somme de micro-activités à caractère scientifique, c'est-à-dire dont la nature et les résultats participent d'une production de connaissances à des états plus ou moins intermédiaires ou finaux, et qui relèvent des méthodologies, des protocoles, des décisions, plus ou moins spécifiques d'une discipline à l'autre ; 2) envisager ces micro-activités comme récurrentes d'une phase à l'autre suggérant plutôt des pratiques concomitantes, itératives, ou tout du moins « en relation ». Or l'une des difficultés récurrentes auxquelles sont confrontés les chercheurs et chercheuses réside justement dans le passage d'un dispositif à l'autre.

6. Dans le sens d'une co-conception des instruments de travail.
7. Voir https://www.ouvrirlascience.fr/category/science_ouverte/.

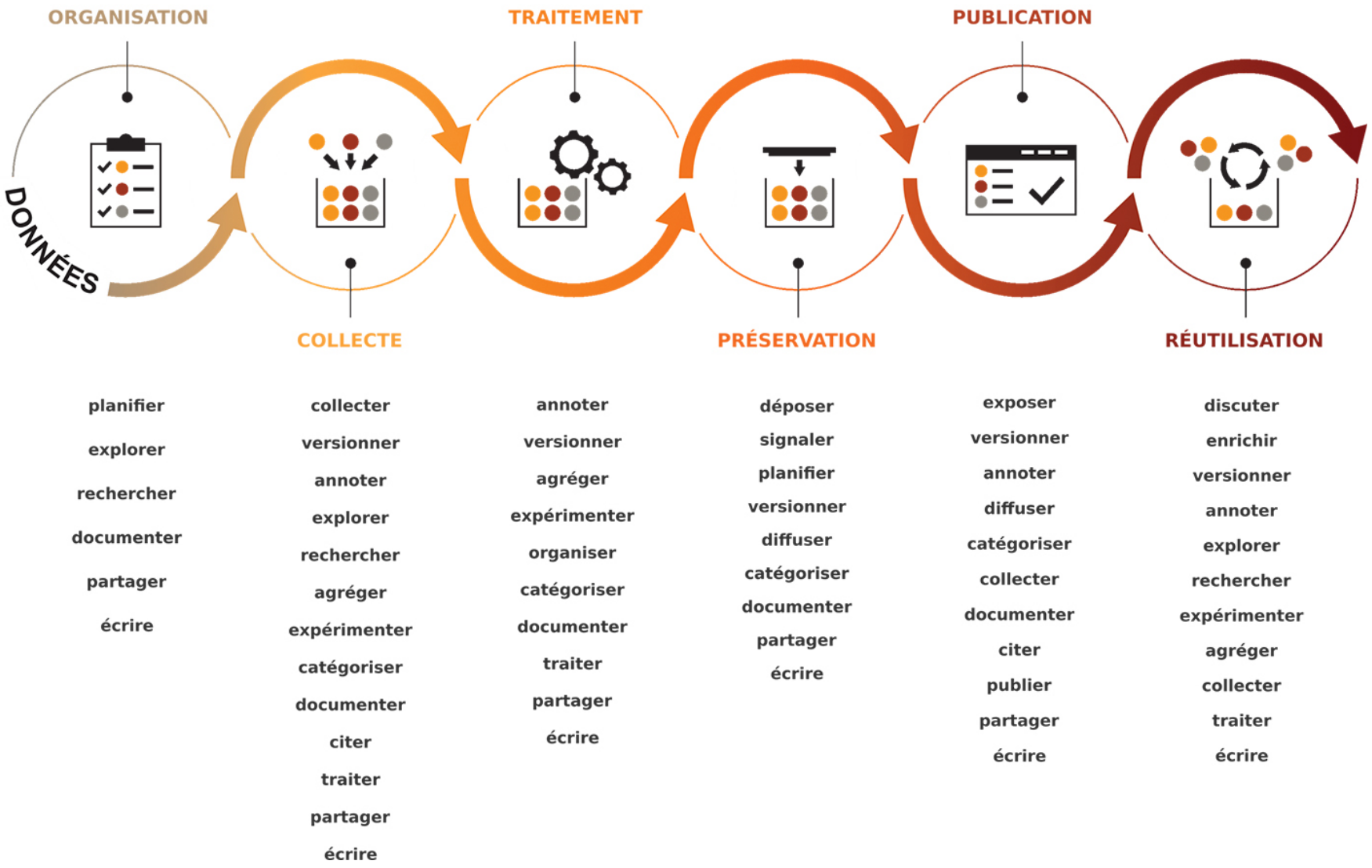


Figure 2. Échantillons de micro-activités de la recherche dans le cycle de la donnée

plateforme	objet	fonctions
Nakala	entrepôt de données	Nakala permet le dépôt des données et leur description (métadonnées) selon différents standards. Un DOI est attribué aux données et celles-ci sont accessibles à travers différents protocoles (API REST, OAI-PMH, visionneuses). Nakala assure à la fois la pérennité, l'accessibilité, la réutilisation des données.
Isidore	agrégateur et moteur de recherche communautaire	Isidore agrège, indexe et enrichit les ressources, données et documents de plusieurs milliers de sources en SHS, proposant un moteur de recherche, des outils de classification et d'enrichissements sémantiques et plusieurs fonctionnalités communautaires associées : partage de recherches, de bibliothèques, suivi d'auteurs et d'autrices, etc.
Stylo	éditeur scientifique	Stylo est un éditeur de texte scientifique en single-source publishing conçu pour les SHS. Stylo est un outil d'écriture et d'édition scientifique basé sur les formats markdown/yaml/bibtex, implémentant des pratiques d'écriture émergentes dans la communauté SHS, et permettant l'export vers les formats de diffusion conformes aux standards de l'édition scientifique en SHS.
Callisto	Jupyter Hub/Lab	Callisto est une instance de Jupyter Hub et Jupyter Lab proposant des kernels pré-installés pour les traitements de données en SHS. Il est destiné à la production collaborative de notebooks, à leur gestion et à leur diffusion.
GitLab	forge logicielle	GitLab Huma-Num est une instance de la forge logicielle GitLab offrant des fonctionnalités communautaires et collaboratives pour la gestion de projets, la production de documents et de code. Ses usages s'étendent à la production, au stockage et au versionning de données, de notebooks, et de toute production de savoirs liés à l'activité de la recherche

Tableau 1. Fonctions des cinq plateformes constitutives de l'écosystème de recherche Huma-Num

Alors même que les cycles se font plus courts et plus itératifs, voire de manière simultanée, l'enjeu pour les équipes de recherche est de maintenir d'une phase à l'autre la cohérence scientifique des données.

9. Ainsi, l'enjeu que le HN Lab cherche à adresser dans cette réflexion ne réside pas dans les dispositifs eux-mêmes, dont les usages peuvent différer d'une communauté utilisatrice à l'autre, mais davantage dans les passerelles et les articulations susceptibles d'être développées entre les dispositifs. Cet horizon nécessite de densifier les interactions entre les services, afin d'assurer la continuité et la cohérence scientifique du processus de recherche d'un espace d'écriture à l'autre. C'est cette densification des échanges qui permettra de tendre vers un écosystème favorisant cette activité distribuée et disséminée, et finalement d'inscrire la recherche dans une dynamique continue et processuelle.

10. Dans notre propos, cinq plateformes composent l'essentiel de cet écosystème en devenir.

11. À titre indicatif, on pourrait tenter de répartir les micro-activités de la recherche à l'articulation de ces cinq plateformes, afin d'illustrer l'idée que le processus

de recherche se compose d'actions précises dont l'exécution est loin de ressembler à un algorithme séquentiel.

12. L'enjeu principal de la conception de cet écosystème réside dans les articulations qui sont ou seront implémentées entre les plateformes. La pertinence et la densité des interactions entre plateformes garantissent la cohérence nécessaire pour faire écosystème. Cette cohérence se joue à la fois sur les protocoles et les formats d'échanges, mais également sur les pratiques qui agiront ces interactions.

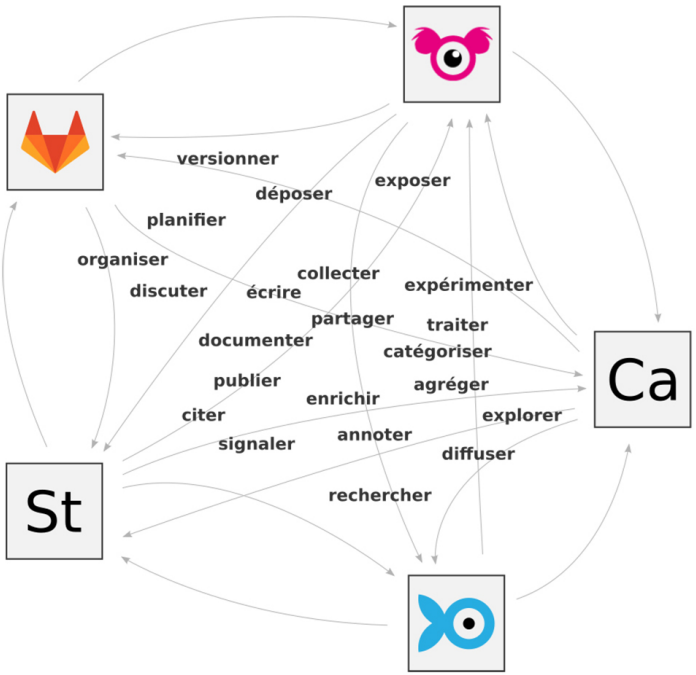


Figure 3. Répartition indicative des micro-activités de la recherche à l'articulation des 5 plateformes considérées

GitLab

13. La plateforme GitLab a vocation à jouer un rôle prépondérant tout au long du cycle de la recherche. GitLab peut en effet intervenir autant dans la gestion des projets (collaboration) que dans la gestion des données, par exemple pour en organiser l'édition, le partage et les traitements. Concernant les données, il est envisagé de développer une fonctionnalité de dépôt simplifié vers Nakala des données produites et stockées sur GitLab. Cette fonctionnalité permettrait notamment de publier des données versionnées tout au long d'un projet. L'identifiant DOI attribué par Nakala à la donnée déposée pourrait ainsi être communiqué et récupéré par le répertoire GitLab associé.

14. Entre Callisto et GitLab, le lien est pratiquement structurel, tant l'espace de travail des utilisateurs et utilisatrices de Callisto n'est pas destiné à être sauvegardé. La bonne pratique consiste alors à stocker et à versionner son travail (données, notebooks, etc.) sur un ou plusieurs répertoires GitLab. C'est également à travers les fonctionnalités de publication et de partage de GitLab que la collaboration au sein d'une équipe ou d'une communauté peut être mise en place pour le traitement et l'analyse de données sur Callisto. On pourrait envisager une pratique similaire pour les sources des documents édités sur Stylo. Bien que la plateforme Stylo propose des fonctionnalités de partage et de collaboration, l'usage de GitLab pour le dépôt et le versionning des sources Stylo ouvrent des possibilités d'édition, de production de formats plus

élaborés que les exports proposés nativement par Stylo. Actuellement, plusieurs éditeurs de revues scientifiques utilisent Stylo pour l'édition et la production des articles, mais le stockage et les traitements finaux avant publication sont déportés sur GitLab qui joue un rôle à la fois de répertoire de travail partagé et d'archivage des sources de référence. Ce scénario est un bon exemple de la convergence de pratiques témoignant d'une maîtrise progressive de l'écosystème et de son maillage fonctionnel.

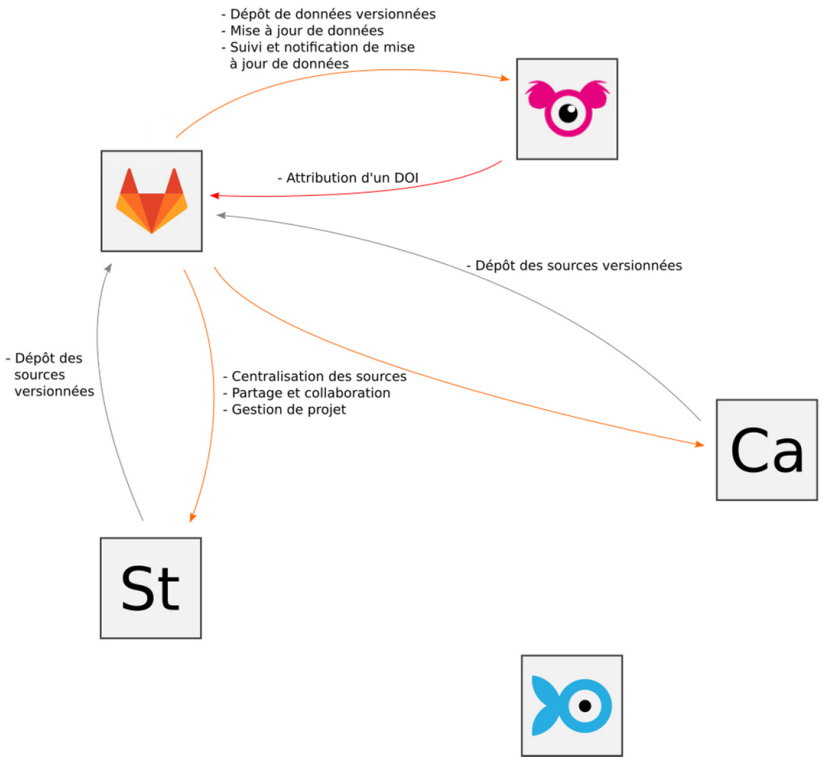


Figure 4. Détail des entrées-sorties pour la plateforme GitLab

Nakala

15. Dans cet écosystème, Nakala assure la fonction de stockage et de publication des données de la recherche, garantissant la pérennité et la citabilité des données. La plateforme est par ailleurs conçue pour favoriser la réutilisation des données, par simple citation (DOI), par leur éditorialisation ou encore par leur consommation dans des chaînes de traitements. La première est assurée par l'attribution d'un DOI pour chaque dépôt de données, auquel s'ajoute un identifiant par fichier déposé. L'éditorialisation est rendue possible par plusieurs visionneuses web permettant d'insérer facilement dans divers documents ou édition HTML une représentation intelligible des données. Enfin, l'API Nakala donne un accès direct et granulaire aux données, servies selon différents formats adaptés aux pratiques de traitements scientifiques des données. Les données de Nakala sont donc susceptibles d'être citées et/ou éditorialisées dans un document Stylo ou encore appelées et analysées à la volée dans un carnet Callisto. On peut noter à ce stade le rôle central des APIs dans l'interconnexion des plateformes. En effet, l'API Rest, caractérisée par les principes d'ouverture et de réutilisation, matérialise notre vision du maillage informationnel.

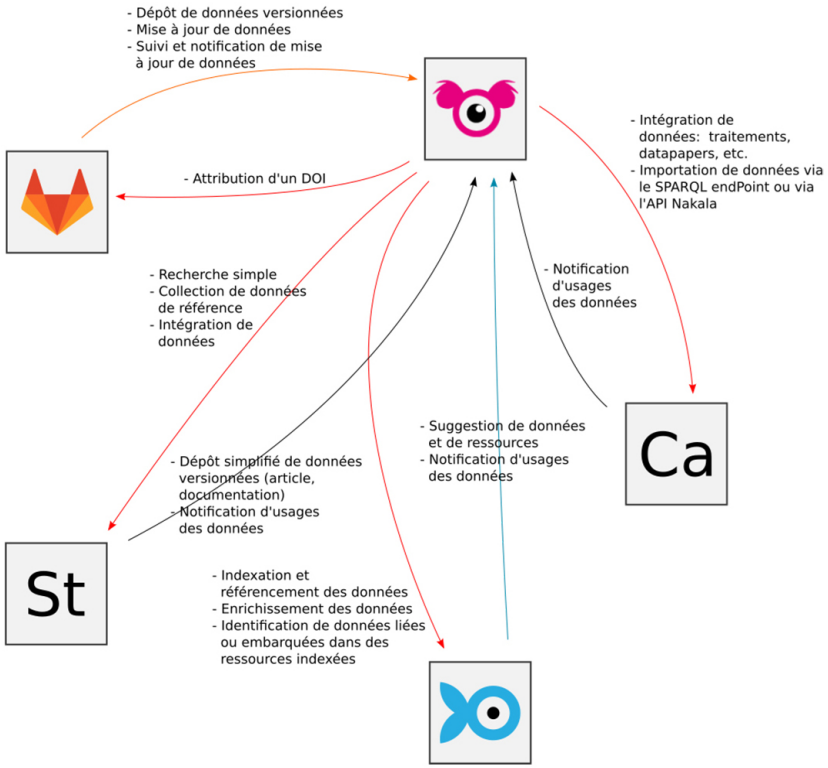


Figure 5. Détail des entrées-sorties pour la plateforme Nakala

16. Si Nakala est l'un des services phares d'Huma-Num, il reste inévitablement associé au moteur de recherche Isidore, car l'approche intégrée de ces deux plateformes a permis de construire une synergie vertueuse pour assister les chercheurs et chercheuses en SHS dans leurs travaux. En revanche, si les interfaces homme-machine des deux services ont été conçues pour être cohérentes entre elles, il n'en reste pas moins que leur complémentarité en termes de partage et de réutilisation des données est en réalité mal comprise et mal exploitée par les utilisateurs et utilisatrices. C'est pourquoi une imbrication plus étroite des deux plateformes, basée sur les principes de la science ouverte, constitue une réelle perspective pour améliorer la maîtrise de la donnée par les communautés SHS. C'est dans cette optique que le programme Huma-Num science ouverte (HNSO)[8] a démarré en février 2021. Le projet s'attache à analyser les deux dispositifs socio-techniques dans leurs dimensions informatique, scientifique et documentaire, afin d'en identifier les interactions et de les renforcer. Cette étude prend la forme d'un ouvrage actuellement en production associé à la réalisation de rapports techniques pour la conception de nouvelles fonctionnalités.

17. De ce point de vue le couple Nakala-Isidore est particulièrement représentatif de la vision proposée dans ce chapitre pour un écosystème de recherche en SHS. Le projet HNSO et son chantier « Interconnexion entre les plateformes Isidore et Nakala » ont déjà mis en évidence deux types d'interactions.

18. La première est indirecte et repose sur l'utilisation de systèmes en commun, en particulier le dispositif d'authentification HumanID[9] qui joue un rôle pivot dans notre écosystème. Si son potentiel d'exploitation

8. Le programme HNSO est financé avec le soutien du Fonds national pour la science ouverte, voir https://hnlab.gitpages.huma-num.fr/blog/projets/.
9. Voir https://humanum.hypotheses.org/5754.

est actuellement encore limité, il a vocation à générer de nouvelles passerelles entre les plateformes. Dans le cadre du projet HNSO, une fonctionnalité émergeant directement de cette authentification partagée consiste à permettre à une personne connectée et disposant de contenus « revendiqués » dans Isidore de recevoir des suggestions automatiques et profilées par disciplines et/ou par thématiques pour l'indexation de ses données lors d'un dépôt dans Nakala. Inversement, il serait possible d'utiliser les données déposées par un chercheur ou une chercheuse dans Nakala pour suggérer dans Isidore des lectures liées sur le plan thématique[10].

19. Les deux plateformes proposent par ailleurs de l'enrichissement par l'intermédiaire de référentiels spécialisés. Lors du dépôt de données dans Nakala, l'utilisateur ou l'utilisatrice peut décrire ses données grâce à la métadonnée « mots-clés ». Celle-ci propose en autocomplétion tous les labels des concepts des référentiels qu'utilise Isidore pour enrichir ses données. En revanche, la fonctionnalité ne conserve actuellement que la chaîne de caractères, le lien avec le référentiel n'existe donc plus. Le projet HNSO propose d'améliorer cette fonctionnalité et renforcer le lien entre les deux plateformes en permettant aux déposants de décrire leurs données avec des référentiels métiers grâce à des propositions d'auto-complétion précisant à quel référentiel le concept choisi appartient, tout en tenant compte des disciplines dans lesquelles le déposant a déjà publié (information récupérée elle aussi via Isidore).

20. Le second type d'interactions concerne l'interconnexion entre les deux plateformes. Celle-ci existe aujourd'hui uniquement par la fonctionnalité de moissonnage des collections Nakala par Isidore au travers du protocole OAI-PMH. Isidore moissonne pour le moment ses collections sur un rythme mensuel, en fonction des sources communiquées par courriel à l'équipe Isidore. Le projet HNSO envisage de fluidifier ce processus, en permettant au déposant de signaler sa collection à Isidore directement depuis l'interface Web de Nakala, pour un moissonnage instantané, avec pour conséquence l'amélioration de la visibilité des données dans le moteur de recherche.

Callisto

21. Basé sur la plateforme Jupyter Hub-Jupyter Lab, Callisto reste à ce jour un prototype dont l'objectif est d'explorer les usages et les besoins spécifiques de la communauté SHS en matière de traitements des données de la recherche. Il s'agit par exemple d'identifier quels kernels (environnement de développement) doivent être mis en place, avec quelles bibliothèques logicielles de traitements. Plus particulièrement, Callisto pourra mettre à disposition une série de templates développés spécifiquement pour les SHS, avec par exemple des carnets prêts à l'emploi pour effectuer certaines tâches

10. Voir https://zenodo.org/record/3991994#.YSzs344zYuU.

génériques[11], ou encore des modèles de réseaux de neurones pré-entraînés sur des corpus spécialisés.

22. Nos observations indiquent que les usages en SHS de ces plateformes se diversifient, intervenant à différents stades de l'activité de recherche : co-développement, pré-traitement et exploration des données, co-écriture (article exécutable), évaluation et validation par les pairs d'une méthodologie ou d'un traitement, publication et diffusion de résultats. Différents usages pédagogiques émergent également, favorisés par la cohabitation d'éléments informatiques de traitement des données et d'éléments discursifs de documentation, d'analyse ou d'interprétation. Cette cohabitation illustre particulièrement la convergence des écritures telles que nous la développons plus en avant dans notre chapitre.

23. L'espace d'écriture et de développement des carnets constitue en effet un nœud où sont réunies la donnée dans sa modélisation mathématique et informatique, la méthode et son implémentation informatique, l'expérimentation et son analyse, jusqu'au discours interprétatif. Le fait que cet espace d'écriture instancie un environnement web n'est pas un hasard, tant le Web et Internet constituent désormais l'environnement de savoirs par excellence, tant pour sa production, sa publication et sa circulation. L'intérêt de Callisto est de déplacer les activités de la recherche nativement dans l'espace du Web, favorisant alors son partage et sa circulation. De fait, l'une des principales nouveautés de ce type de plateforme par rapport à des méthodes plus traditionnelles de développement réside dans une dynamique de collaboration inédite, caractérisée par cet espace d'écriture où cohabitent et collaborent les différents membres d'une équipe. L'intégration de Callisto dans l'écosystème encourage encore davantage les approches collectives du travail de recherche.

24. Les interactions de Callisto avec les autres plateformes sont particulièrement évidentes pour les données entrantes (Nakala, GitLab, Isidore) dans Callisto. Ces interactions peuvent intervenir au niveau du carnet, par exemple lorsqu'une fonction extrait de l'entrepôt Nakala un jeu de données pour l'exploiter. Elles peuvent intervenir également au niveau des plateformes elles-mêmes, par exemple avec l'intégration dans Callisto d'un module GitLab dédié bénéficiant de l'authentification partagée HumanID. L'interconnexion des deux services apporte une solution de stockage et de versionnage des carnets. En effet, l'espace de travail ouvert dans Callisto n'est pas conçu pour être pérenne ni accessible en continu. En reprenant l'analogie de la paillasse de laboratoire, l'espace de travail personnel dans Callisto doit être régulièrement rangé et nettoyé pour éviter toute perte de données. Le répertoire GitLab intervient alors comme l'armoire où l'on range ses ustensiles et ses artefacts. Une telle bonne pratique ne relève pas que de la sauvegarde de son travail. Cette maintenance sur GitLab en tant que plateforme communautaire permet aussi de rendre son travail accessible à d'autres. Son association avec

11. C'est le modèle développé par le projet ModOAP dirigé par Julien Schuh dans le cadre du labex Les passés dans le présent.

l'environnement Callisto ouvre les carnets Callisto au partage, à la collaboration et à la validation scientifique.

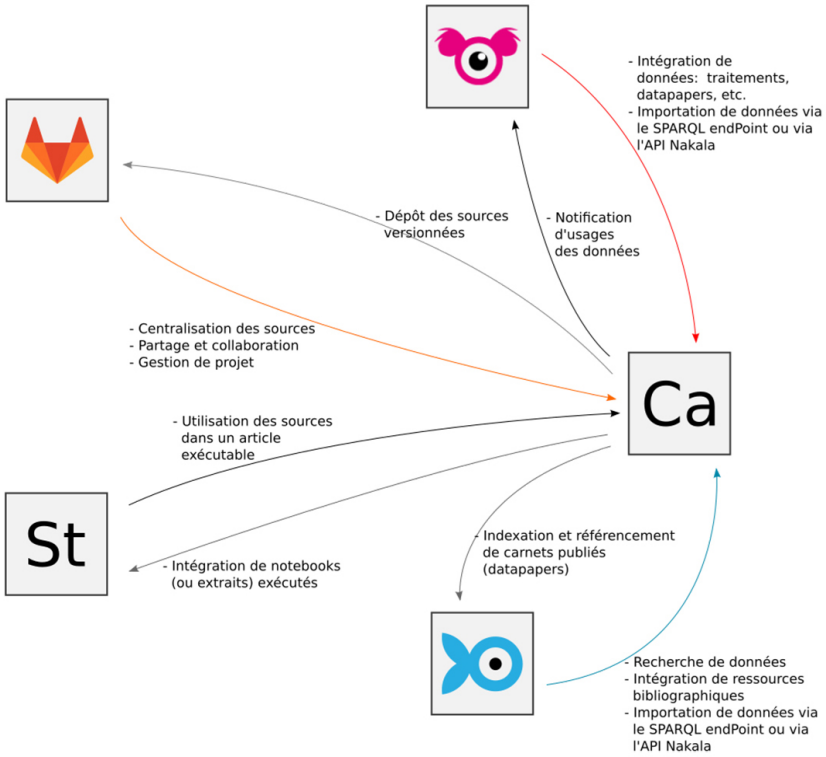


Figure 6. Détail des entrées-sorties pour la plateforme Callisto

25. Les carnets Callisto peuvent exploiter directement les données de Nakala et Isidore, en profitant de leur API, et dans le cas d'Isidore de son triple store. En retour, les carnets développés sur Callisto ont également vocation à alimenter l'ensemble de l'écosystème : dépôts sur GitLab de données produites dans Callisto, notifications et métriques d'usages de données vers Nakala, éditorialisation de carnets ou de cellules de carnets[12] dans un document Stylo, ou encore indexation de carnets publiés dans Isidore.

Stylo

26. Stylo a été conçu comme un outil d'écriture et d'édition de documents et d'articles scientifiques. Plusieurs enjeux ont motivé la réalisation de cette plateforme, notamment la mise en place d'une chaîne de production de l'écrit vertueuse sur le plan scientifique. Là où les chaînes éditoriales numériques traditionnelles engendraient des ruptures de sens entre l'auteur, l'éditeur et le diffuseur, Stylo assure au contraire la continuité des données et de la structure des documents. La plateforme d'édition implémente également l'état de l'art en matière de formats et de standards ouverts. De ce point de vue, la philosophie de Stylo consiste à faciliter l'appropriation auprès des chercheurs et chercheuses et des éditeurs et éditrices de bonnes pratiques, considérant ces dernières comme une marche non négligeable dans l'accès à la littératie numérique.

27. De nombreuses passerelles vers et depuis Stylo sont susceptibles de densifier le réseau d'interconnexions de

12. Dans un carnet Jupyter, une cellule désigne un bloc de code ou un bloc de texte.

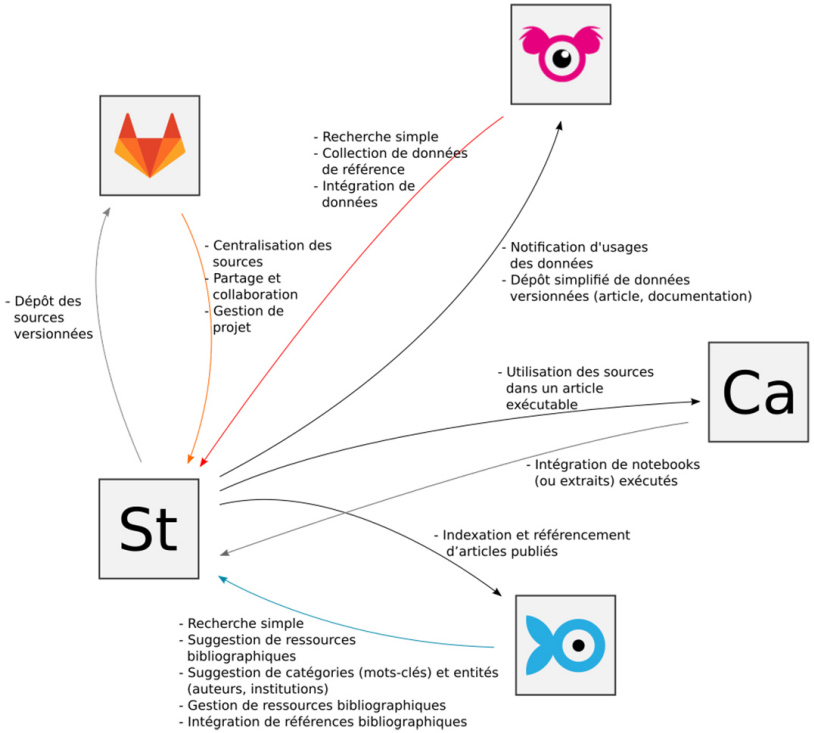


Figure 7. Détail des entrées-sorties pour la plateforme Stylo

l'écosystème. Selon les passerelles envisagées, les usages de Stylo tendront à se diversifier, de l'article scientifique jusqu'aux *data papers*, en passant par la rédaction de tout fragment textuel tirant une pertinence scientifique de son interconnexion à l'écosystème.

28. Sur un modèle similaire au module Zotero qui permet déjà de récupérer et d'exploiter les ressources bibliographiques d'une collection Zotero en ligne, un module Isidore permettra de lister et d'exploiter des ressources bibliographiques venant d'Isidore. Un premier scénario en mode non connecté consiste à exploiter l'API Isidore pour lancer des recherches puis à sélectionner des ressources, ou encore à ajouter manuellement dans Stylo les Handle de ressources identifiées sur isidore.science. Un second scénario en mode connecté, c'est-à-dire bénéficiant de l'authentification partagée HumanID, consiste pour l'utilisateur et l'utilisatrice de Stylo à récupérer une collection Isidore personnelle comme source bibliographique pour la rédaction de son document. Ces deux scénarios sont également envisagés pour l'interconnexion de Nakala et de Stylo pour la citation et l'éditorialisation de données déposées dans Nakala, que ces données soient publiques (mode non connecté) ou privées (mode connecté). En intégrant un module Nakala facilitant la réutilisation des données dans le corps de texte, Stylo peut effectivement étendre son champ d'usages à la rédaction et à l'édition de *data papers*. Cette extension d'usage ne présente pas en soi de complexité technique ou pratique particulière. En effet, l'écriture d'article dans un environnement nativement web s'enrichit très naturellement de tout fragment de ressources déjà présent sur le Web. De ce point de vue, les modules de visualisation de données déjà prévus par Nakala révèlent toute la cohérence d'un écosystème intégré. Le verrou sera davantage éditorial selon les formats imposés lors de la diffusion d'articles intégrant des données éditorialisées. Une attention particulière doit ainsi être portée sur les modalités techniques et éditoriales de diffusion des *data papers*.

29. Par ailleurs, Stylo a aussi vocation à s'immiscer à l'intérieur des autres plateformes pour écritures diverses, associées par exemple à une donnée ou une collection de données (Nakala) pour leur documentation, ou à une référence bibliographique Isidore pour l'écriture d'une note de lecture, ou encore à une collection de références bibliographiques pour lancer l'écriture d'un article. Dans ce cas, le document Stylo se verrait distribué sur les plateformes tierces pour des usages situés (documenter une donnée, partager une feuille de route, etc.) ou accessible dans Stylo pour une pratique éditoriale plus complète.

Isidore

30. Isidore est un moteur de recherche permettant de découvrir et de trouver des publications, des données numériques et des profils de chercheurs et de chercheuses en SHS du monde entier. Il offre une recherche dans le texte intégral de plusieurs millions de documents (articles, thèses et mémoires, rapports, jeux de données, pages Web, notices de bases de données, description de fonds d'archives, signalements d'événements scientifiques, etc.). De plus, Isidore relie entre eux ces documents en les enrichissant de concepts scientifiques eux-mêmes issus des travaux des communautés de recherche en SHS. Il est accessible sur le Web sur le portail isidore.science mais également au travers d'une API et d'une exposition dans le Linked Open Data via un SPARQL endpoint.

31. Les API de recherche et de suggestion de publications et de données d'Isidore positionnent la plateforme dans un rôle de hub permettant la dissémination d'informations et de données pour les applications telles que Nakala,

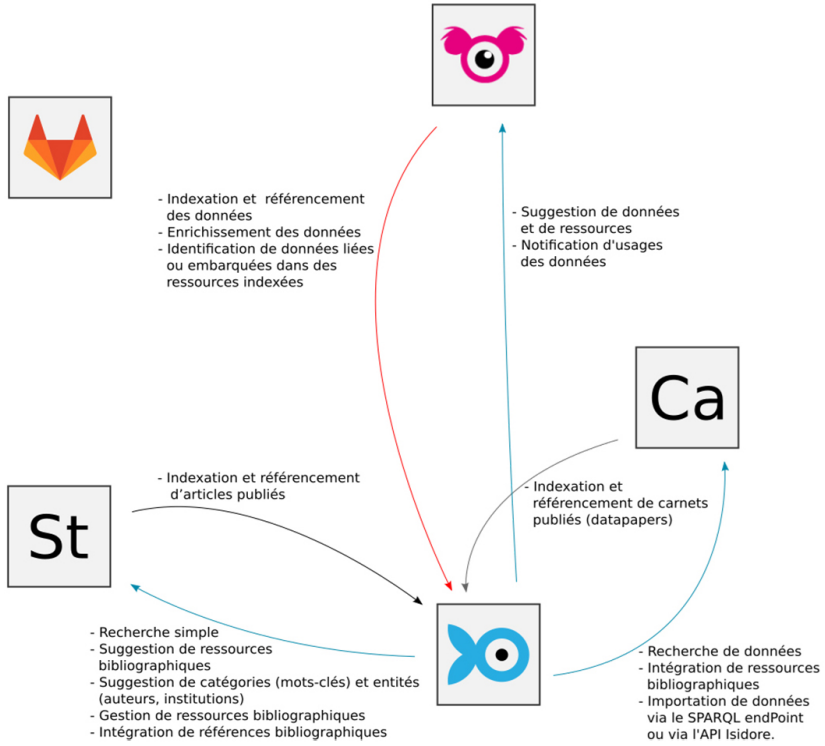


Figure 8. Détail des entrées-sorties pour la plateforme Isidore

Stylo ou Callisto. Les capacités d'ingestion de données et de métadonnées d'Isidore permettraient d'indexer des documents complexes tels que des dépôts GitHub (par

exemple à partir d'un readme.md structuré) ou tout autre document pouvant être décrit par des métadonnées.

32. En aval de l'indexation et de l'enrichissement sémantique à l'aide de référentiels scientifiques, Isidore pourrait être interrogé lors de la rédaction d'articles dans Stylo pour suggérer des références bibliographiques pertinentes, sur le modèle du projet Isidore Suggestion (Pouyllau, 2016).

Discussion

Convergence des écritures

33. L'écosystème ainsi anticipé dans ses multiples interactions fait émerger l'écriture comme une activité centrale tout au long du processus de recherche. Pour bien comprendre cette affirmation, réévaluons tout d'abord la notion même d'écriture scientifique. Cette dernière est généralement associée à la communication scientifique, en fin de cycle, lorsque les résultats d'une recherche sont considérés par l'équipe ou le chercheur comme publiable. L'écriture relève alors de la production d'un discours et d'une forme tous deux élaborés selon des normes et des attendus tant scientifiques qu'institutionnels (Pontille, 2007 ; Dacos et Mounier, 2010). Si les visées d'un article peuvent varier, la teneur générale, notamment en SHS, introduit une démarche critique et interprétative sur la base de résultats de recherche obtenus, que ceux-ci soient issus d'un terrain, d'un corpus, d'une enquête, ou de tout autre « objet de recherche ». L'article scientifique se caractérise ainsi par un certain nombre de marqueurs formels (longueur, structure, etc.), scientifiques (appareil critique, bibliographie, auctorialité) ou encore institutionnels (institution des auteurs et autrices, revue, éditeur, diffuseur, etc.).

34. Or, dans l'activité scientifique suggérée par l'écosystème décrit, l'écriture intervient bien avant la rédaction d'un article. Elle se niche par exemple dans la production d'un ticket explicitant une méthode à employer, dans la documentation de l'encodage éditorial d'un document numérisé, dans la documentation d'un commit améliorant un algorithme de traitement, etc. Pour aller plus loin, l'écriture est déjà présente dans la pratique scripturale du code informatique de ce même algorithme. De même, l'encodage dans un format TEI ou autre format de balisage relève également de l'inscription selon des modalités spécifiques.

35. Ainsi, notre conception de l'écriture fait davantage appel à une activité continue, itérative et incrémentale, distribuée parmi les différents espaces d'écriture qui caractérisent les activités unitaires de la recherche. Les statuts de ces écrits ne sont pas les mêmes, tant sur le plan informatique que sur le plan scientifique, mais tous participent d'une production fragmentée et modulaire de connaissances dès lors que ces écrits s'inscrivent dans et entre les plateformes composant l'écosystème. À travers ce dernier, nous faisons l'hypothèse d'un espace scriptural pluriel de production et de diffusion des connaissances. Les *notebooks*, de type Jupyter ou

R Markdown par exemple, sont emblématiques de cet espace pluriel puisqu'ils font cohabiter dans un même fichier et une même interface les données, leurs traitements et l'analyse scientifique (discursive) de ces traitements. À l'échelle de l'écosystème, la pluralité et la cohabitation s'expriment sur les différentes plateformes et se rejoignent au fil des liens tissés entre elles (par exemple, la documentation sur Stylo d'une donnée déposée dans Nakala) ou en elles (par exemple, l'écriture dans Stylo d'un *data paper* intégrant une donnée déposée dans Nakala).

36. Cette pluralité est aussi le lieu d'une convergence. Si l'on regarde l'écosystème du point de vue des pratiques d'écriture, on remarque en effet la possibilité d'un espace interdisciplinaire où les pratiques et donc les positions épistémiques convergent. Il faudrait probablement caractériser précisément cette convergence, mais on peut d'ores et déjà identifier comme un signal faible notable l'usage récurrent du format Markdown pour l'écriture et l'édition d'éléments aux fonctions scientifiques diverses : ticket ou encore wiki dans GitLab, cellule de notebook dans Callisto, article scientifique dans Stylo, etc. Les « flavors » du format ne sont pas les mêmes d'une plateforme à l'autre, mais ce n'est pas un hasard si les principes originels du format Markdown tendent à s'imposer comme format d'écriture scientifique (Mpondo-Dicka, 2020). En effet, le Markdown étant nativement interprété en HTML, il permet d'éditer et de publier du contenu directement dans le Web, devenu de fait l'espace universel de connaissances[13].

37. Cette convergence de format est aussi une convergence de pratiques qui ensemble entrouvrent la porte à une approche progressive de la littératie numérique. Les bases ou les bonnes pratiques acquises pour une plateforme donnée constituent déjà un point d'entrée aux autres plateformes pour de nouveaux usages.

Ouverture de la recherche

38. Une autre tendance transparaît de cet écosystème ainsi décrit, dans le sillon tracé par les directives de la science ouverte et des principes FAIR. Si le libre accès constitue la pierre angulaire de la science ouverte, il concernait dans un premier temps l'accessibilité des publications scientifiques, puis plus récemment des données de la recherche. Les plateformes Isidore et Nakala ont spécifiquement été conçues dans cette optique d'ouverture, d'accès et de découvrabilité des données et des publications. Mais la vision que nous proposons avec cet écosystème projette l'ouverture de la recherche un pas plus loin en considérant l'ouverture des protocoles et des méthodologies de recherche, voire, selon les modalités d'ouverture et de publication envisagées, du processus même de la recherche. On peut particulièrement

13. Précisons que le Markdown n'a pas vocation à remplacer les formats HTML ou XML sur lesquels sont établis la plupart des formats de diffusion, mais il a vocation à devenir le format de rédaction, voire d'édition lorsqu'il est pleinement exploité.

observer cette nouvelle ouverture sur des plateformes collaboratives telles que les forges logicielles, GitHub et GitLab dont plusieurs instances sont déployées par les institutions de la recherche. Conçues pour synchroniser les contributions à un code informatique, ces forges ont rapidement accueilli de nouveaux usages dans le monde de la recherche, et sont désormais utilisées pour l'édition collaborative de données, pour l'écriture d'ouvrages et d'écrits scientifiques, ou encore désormais pour faire de l'édition et de la publication continue. Ces espaces sont devenus à la fois des lieux de production et des lieux de publication de la recherche en train de se faire. On y observe un véritable déplacement des modes de communication scientifique, vers des formats intermédiaires et directement opérationnels.

39. Que ce soit à des fins de transparence et de traçabilité du raisonnement (Granger, 2020), ou à des fins de reproductibilité de la recherche (Rey-Coyrehourcq *et al.*, 2017), ce partage du processus même de la recherche vient mobiliser les registres de la documentation et de la pédagogie. Ainsi, ces pratiques que l'on retrouve dans des initiatives de plus en plus structurées, comme les sites programminghistorian.org ou rzine.fr, constituent un fort levier de littératie numérique dans les communautés SHS.

40. Notons que notre proposition ne se positionne pas comme une alternative à la publication scientifique classique, outillée au niveau national par des infrastructures dédiées, mais comme une solution générique de partage de connaissances. S'il est vrai que les modalités de la communication scientifique s'en trouvent élargies, le partage du processus s'inscrit à la fois dans le processus et à sa périphérie, puisqu'il est porteur de dynamiques collaboratives au sein et au-delà des équipes de recherche, vers la communauté de recherche.

Intégration modulaire

41. L'écosystème proposé oriente la définition d'infrastructure de recherche vers un ensemble de dispositifs socio-techniques, définis et conçus par agrégation et articulation de briques logicielles adaptées aux besoins d'une recherche reproductible, réutilisable et correspondant aux pratiques des communautés SHS. De ce point de vue, l'infrastructure ne peut plus être envisagée comme une simple fourniture d'outils isolés (qu'ils soient sur mesure, uniques, ou déjà standardisés). Elle doit en effet être conçue comme un assemblage modulaire définissant, entre les outils et à partir des pratiques de recherche, des parcours et des chemins ré-empruntables et scientifiquement critiquables. Il s'agit de concevoir les outils à travers leur « urbanisation », c'est-à-dire à travers leur implantation au croisement de pratiques et à travers les interconnexions au reste de l'écosystème de recherche. C'est un champ de possibles qui s'ouvre ainsi.

42. Là où les GAFAM ont tendance à mettre en place des écosystèmes intégrés opaques caractérisés par l'incorporation des services dans une même plateforme, l'écosystème de recherche proposé tend au contraire à une

intégration modulaire d'outils et de services interopérables, garantissant la transparence dans les protocoles d'échanges. Cette approche se révèle en effet plus cohérente et plus pérenne en termes de données, mais aussi plus accueillante aux pratiques déjà existantes.

43. La modularité, l'interopérabilité et la transparence qui caractérisent une telle intégration d'outils tendent à favoriser chez les chercheurs et les chercheuses une meilleure littératie numérique, que l'on comprenne d'ailleurs celle-ci comme un éthos culturel et social dans l'environnement numérique, ou comme la somme des compétences nécessaires pour y lire, écrire, structurer ou encore publier, et pour la recherche pour en classer, traiter, analyser, catégoriser les données.

44. En effet, ces trois qualités inscrivent d'une part l'activité de recherche et d'écriture dans un processus collectif en favorisant les bonnes pratiques en termes de partage et de collaboration. D'autre part, elles rendent visibles les flux de données tout au long du processus de recherche, incitant les chercheurs et les chercheuses à se saisir des enjeux sur les formats qu'ils manipulent. Elles créent ainsi les conditions pour 1) une appropriation de bonnes pratiques et 2) une autonomisation de l'utilisateur et de l'utilisatrice vis-à-vis de l'outil, et ouvre la voie au développement d'une pratique experte de l'« écriture scientifique ». C'est en cela que nous devons comprendre l'écriture scientifique comme la somme des pratiques hétérogènes qui constituent l'activité de recherche, et comme le point focal de la littératie numérique.

Pour consulter les données mobilisées dans le chapitre, voir :
CSV « Synthèse des articulations entre les cinq composants de l'écosystème » [doi:10.34847/nkl.54afy3ed/7befd28456ce1e62db46a68ffbabffd40d631d95] - Creative Commons Attribution Share Alike 4.0 International (CC-BY-SA-4.0)

HTR-United : un écosystème pour une approche mutualisée de la transcription automatique des écritures manuscrites

Alix Chagué, Thibault Clérice et Laurent Romary

Introduction

1. Depuis quelques années, les projets en humanités numériques intègrent des tâches de transcription automatique d'écritures manuscrites pour l'acquisition des corpus, confirmant le transfert de cette technologie du domaine expérimental de la vision par ordinateur vers le grand public. En témoigne le développement de logiciels conviviaux, libres ou propriétaires, proposant des solutions quasi clefs en main, tels que Transkribus [Kahle *et al.*, 2017], eScriptorium [Stökl Ben Ezra, 2021] ou encore Arkindex [Teklia, 2021]. Parmi les projets ayant eu recours à ces logiciels, on peut citer Himanis [Stutzmann *et al.*, 2017], Ffl [Massot *et al.*, 2019], Horae [Boillet *et al.*, 2019], Time US [Chagué *et al.*, 2019], MaRITEM [Mariotti, 2020], Lectaurep [Chagué *et al.*, 2020]. On pourrait en déduire que n'importe qui peut désormais se lancer dans un projet de reconnaissance automatique d'écritures manuscrites, mais il reste en réalité de nombreux points de blocage. Ainsi, bien qu'elles soient à portée de main, les plateformes techniques implémentant des solutions de transcription automatique ne sont pas encore en mesure de traiter toutes les formes d'écritures manuscrites et nécessitent de grandes quantités de données pour cela. Produire ces données a un coût que la mutualisation des efforts peut atténuer.

2. Nous présentons dans ce chapitre un écosystème nommé HTR-United [HTR-United *et al.*, 2020/2021] facilitant la mise en commun de la vérité de terrain. Cette solution propose un modèle opérationnel susceptible d'offrir un cadre pour la construction de *data papers* pour l'HTR, voire les prémices d'une standardisation pour ce genre de publication. Nous commençons par rappeler le fonctionnement de la transcription automatique ainsi que ses limites actuelles. Nous démontrons ensuite l'importance stratégique de décloisonner les données issues des activités préparatoires à l'HTR avant de présenter la solution mise en œuvre par l'intermédiaire de l'écosystème HTR-United. Cet écosystème est construit dans une logique minimaliste et s'appuie sur la plateforme GitHub ; nous en détaillons le fonctionnement et la structure. Enfin, nous revenons sur l'importance de mettre en place un contrôle qualité sur les données publiées et présentons les outils intégrés dans l'écosystème HTR-United pour aider à cela.

Principes de la transcription automatique

3. La reconnaissance des écritures manuscrites, que l'on appelle aussi HTR (*Handwritten Text Recognition*), est un procédé informatique qui vise à obtenir un équivalent de texte numérique à partir de l'image d'un document physique comportant du texte manuscrit. Ce traitement est décomposé en trois tâches (figure 1) dont deux (1, 3) sont indispensables : on commence (1) par localiser l'emplacement du texte sur l'image de manière à produire un ensemble de coordonnées (segmentation) ; puis (2) en fonction des logiciels et des besoins, on peut déterminer automatiquement l'organisation logique de chaque segment par rapport aux autres et par rapport à la page (analyse de la mise en page) ; enfin, (3) on reconnaît les lettres et les mots tracés dans chaque portion de l'image définie par les coordonnées d'un segment (transcription).

4. Ces tâches relèvent du domaine de l'apprentissage profond, il est donc nécessaire d'entraîner, pour chacune d'entre elles, des modèles à partir de données d'exemple. Ce sont ces exemples que l'on appelle la vérité de terrain : des ensembles de données annotées de manière à fournir au modèle des paires composées d'une part d'une image ou d'une portion d'image (entrée) et d'autre part de l'annotation attendue (sortie). Celle-ci peut être des coordonnées dans le cas de la segmentation ou un ensemble de caractères dans celui de la transcription. Les performances des modèles dépendent de l'efficacité de l'architecture neuronale mise en place, mais aussi de la qualité et de la quantité de vérité de terrain fournies lors de l'apprentissage.

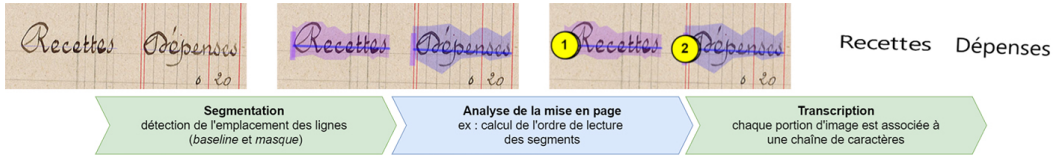


Figure 1. Schématisation des étapes de traitement impliquées dans l'HTR

5. De nombreux facteurs font que la tâche de transcription constitue encore un défi [Stokes *et al.*, 2021]. On peut citer la très grande variation dans la formation des lettres (variation intra et inter-classe), la forte présence de bruit et d'accidents sur les pages manuscrites, l'impossibilité de s'appuyer sur une segmentation à l'échelle des caractères ou encore la présence de graphèmes et de systèmes d'abréviations propres à chaque personne (figure 2). S'y ajoute la difficulté pour les annotateurs et annotatrices de se mettre d'accord sur les pratiques de transcription, notamment la manière de traiter les variations graphétiques [Stutzmann, 2011] ou les abréviations.

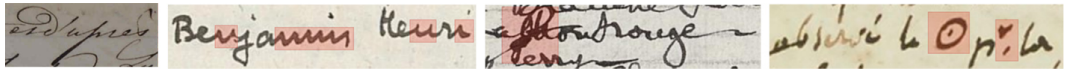


Figure 2. Exemples illustrant les principales difficultés rencontrées pour le traitement de textes manuscrits

6. Malgré ces défis, il est déjà actuellement possible d'obtenir des modèles produisant des transcriptions à 95 % réussies [Pinche, 2021]. Pour produire de tels modèles, il existe deux approches (figure 3). La première configuration s'apparente à un démarrage à froid : on part de zéro et on entraîne un modèle en fournissant de la vérité de terrain à un moteur de transcription et en définissant une architecture neuronale et des hyper paramètres. Dans l'autre cas de figure, on s'appuie sur un modèle préalable que l'on affine. C'est-à-dire que l'on fournit au moteur de transcription un modèle de base plus ou moins performant et dont on reprend les paramètres pour lancer un nouvel entraînement basé sur des exemples plus ou moins similaires à ceux qui ont permis l'entraînement du modèle initial. Cette deuxième approche présente des avantages, parmi lesquels un important gain d'efficacité : en s'appuyant sur les acquis préalables d'un modèle, on a besoin d'une moindre quantité de vérité de terrain pour obtenir de bonnes, voire de meilleures performances. Cela signifie qu'au lieu de devoir transcrire manuellement une centaine de pages issues d'un corpus nouveau, on peut se contenter d'une trentaine de pages tout en parvenant *in fine* aux mêmes performances [Reul *et al.*, 2021].

7. Dans les deux cas toutefois, en matière de transcription automatique pour les écritures manuscrites, il est rare de pouvoir se passer d'un entraînement sur des exemples correspondant au corpus à traiter, à l'inverse de la transcription automatique de l'imprimé ou bien de la segmentation, où les modèles disponibles sont déjà suffisamment performants. Cela signifie qu'il est presque toujours nécessaire de commencer par la transcription manuelle d'un échantillon du corpus.

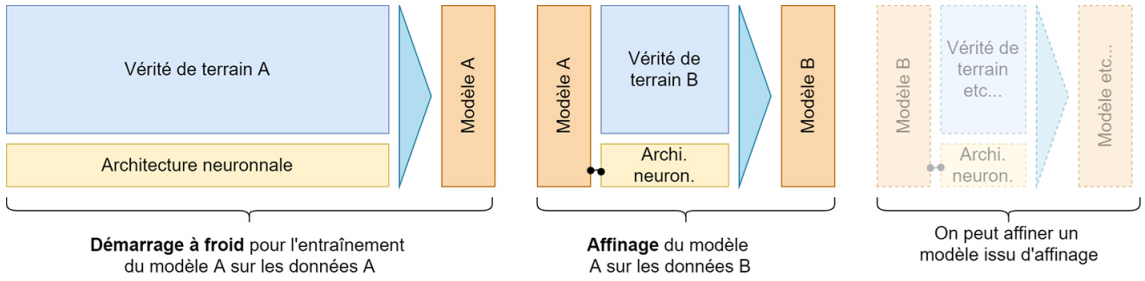


Figure 3. Les deux configurations principales pour l'entraînement de modèles d'HTR ne demandent pas la même quantité de données

Décloisonner les ressources, partager les données

8. Produire de la vérité de terrain de qualité suppose de posséder une connaissance minimale des environnements de transcription automatique, de mettre à plat l'ensemble des règles de transcription permettant d'obtenir une sortie de texte correspondant aux attentes, et surtout de posséder les moyens en temps et financiers de produire cette vérité de terrain. Il faut alors s'assurer de la disponibilité de personnes capables de lire les écrits du corpus, qu'elles disposent d'une compréhension suffisante des documents et des enjeux du projet, et qu'elles soient capables de contrôler

la qualité de la transcription par rapport aux règles fixées en amont. Il est rare qu'un projet en humanités numériques souhaitant recourir à la transcription automatique possède toutes ces ressources.

9. Un écueil guette ces projets, celui d'échouer à obtenir un modèle de transcription efficace et d'abandonner alors l'ambition initiale d'automatiser cette tâche. Le risque est de dédier un temps considérable à produire une transcription qui, d'échantillon de vérité de terrain, finit par devenir le corpus final. Si on prend un peu de hauteur, c'est aussi une perte de ressources pour la communauté des sciences humaines, car, à l'heure actuelle, la vérité de terrain n'est souvent produite que pour servir les finalités d'un projet précis. Il s'opère donc un cloisonnement des ressources où chaque projet part de zéro et doit produire, ou tenter de produire, sa propre vérité de terrain, comme si rien n'avait été transcrit avant.

10. Plusieurs facteurs expliquent le cloisonnement de ces données. En tout premier lieu, on peut considérer que les modèles de transcription (ou de segmentation) sont la partie la plus spectaculaire d'un projet de transcription automatique, car ils permettent à n'importe qui, quelle que soit sa familiarité avec les principes de l'apprentissage profond, de produire immédiatement une sortie textuelle plus ou moins correcte. En plus, il est possible de communiquer les performances théoriques d'un modèle, car elles peuvent être exprimées sous la forme de scores faciles à comprendre : on mesure notamment le taux d'erreur par caractère (CER pour *character error rate*) et le taux d'erreur par mot (WER pour *word error rate*). Comme les logiciels d'HTR permettent aux utilisateurs et utilisatrices de partager des modèles, il semble plus évident de baser sa communication sur le(s) modèle(s) entraîné(s) et employé(s) ainsi que sur la transcription finale obtenue (où les transcriptions manuelles sont souvent mêlées aux transcriptions obtenues automatiquement). À l'inverse, il est peut-être difficile de partager la vérité de terrain : outre parfois l'absence de fonctionnalités adéquates dans les logiciels, les droits sur les images peuvent restreindre les possibilités de partage, les transcriptions peuvent faire l'objet d'embargo et les compétences et capacités de calcul sont encore rares pour (re)produire des modèles à partir de ces données.

11. Lorsqu'elles sont publiées en tant que vérité de terrain, les données d'entraînement sont peu visibles, car il n'existe aujourd'hui, au niveau national comme international, ni entrepôt ni modèle de description dédiés. Pour les personnes susceptibles de réutiliser ces données, il n'est alors pas possible de filtrer rapidement les jeux correspondant à leurs projets. Par exemple, sur un entrepôt généraliste comme Zenodo, on peut chercher des jeux de données liés à l'HTR grâce aux mots-clefs associés aux dépôts, mais il n'est pas garanti que leur description suffise pour évaluer leur qualité. Les autres entrepôts généralistes, y compris Nakala, souffrent de ce même problème : les données HTR y côtoient des données diverses (modèles 3d, éditions de textes, données géographiques, transcriptions d'entretiens, etc.) et, à ce titre, leur recherche est difficile. Quelques tentatives d'énumération

de jeux de données pour l'OCR et/ou l'HTR existent[1], mais elles sont imparfaites à plusieurs titres : aucune ne propose des descriptions suffisamment complètes, la cohérence des jeux de données n'est pas garantie, les critères de sélection ne sont pas explicites ; et elles reposent sur une veille de la part de leurs créateurs.

12. Pourtant s'appuyer sur la vérité de terrain plutôt que sur un modèle présente plusieurs avantages. Ils sont liés en tout premier lieu au fait qu'il est impossible de transposer un modèle d'un moteur de transcription à un autre. En effet, chaque solution d'HTR est basée sur un écosystème de développement pour l'apprentissage machine[2] qui vient avec ses propres spécifications, et l'export et l'enregistrement des modèles dépend des choix de développement propre à chaque moteur. Certains logiciels grand public ne permettent même pas à l'utilisateur ou à l'utilisatrice d'exporter le modèle entraîné : celui-ci n'est alors disponible que par l'intermédiaire du logiciel qui l'a produit et du serveur qui l'héberge. Cette captivité des modèles rend les utilisateurs et utilisatrices, ainsi que leurs projets, vulnérables à l'arrêt des développements des logiciels ou de leurs dépendances, et plus généralement face aux aléas informatiques. À l'inverse, les données d'entraînement s'avèrent plus souples, notamment du fait qu'il existe des standards ouverts comme XML ALTO [ALTO 4.2, 2020] et XML PAGE [Pletschacher et Antonacopoulos, 2010], que la plupart des moteurs d'HTR implémentent, pour enregistrer le résultat des différentes étapes de transcription. Un fichier XML étant essentiellement un fichier textuel, il est aisé de le lire ou de le modifier. Il est donc possible d'exporter les données produites à l'aide d'un logiciel de transcription, de les modifier au besoin, et de les réinjecter dans un autre logiciel de transcription.

13. Dans le cadre de la science ouverte, l'enjeu de la publication des transcriptions finales est certes compris, de même que celui de publier les modèles lorsque cela est possible, mais il manque un réflexe de publier la vérité de terrain en tant que vérité de terrain et non pas en tant que transcription. En fait, cela est d'autant plus dommageable que le fait de pouvoir accéder à la vérité de terrain permet de comprendre quelles ont été les pratiques de transcription conduisant à un modèle, d'en comprendre les résultats et même de reproduire l'entraînement du modèle[3].

14. Outre ces aspects de portabilité des données, il nous faut mentionner la plasticité de la vérité de terrain : il est impossible de fusionner des modèles de transcription, alors qu'on peut assembler, diviser, croiser différents jeux de vérité de terrain pour en recomposer un nouveau. De même, on peut modifier les exemples de transcription qu'ils contiennent de manière à rendre des

1. Par exemple, la liste des jeux de données d'OCR recensés par Awesome-OCR (Konstantin Baierer) sur GitHub (https://github.com/kba/awesome-ocr#datasets) ou encore le travail de Clemens Neudecker (https://cneud.github.io/ocr-gt/) (liens consultés le 12/02/2022).
2. Pour les systèmes basés sur Python, on peut citer PyTorch (https://pytorch.org), Tensorflow (https://www.tensorflow.org) ou encore DyNet (https://github.com/clab/dynet).
3. À condition que l'ensemble des paramètres de l'entraînement ait été documenté.

jeux compatibles entre eux, ou pour obtenir un modèle dont la sortie correspond à nos besoins. On comprend alors qu'accéder à la vérité de terrain d'autres projets permet à coup sûr d'éviter un démarrage à froid : grâce à ces données, on peut créer son propre modèle pré-entraîné pour basculer dans un scénario d'affinage, ou bien augmenter rapidement l'importance matérielle de sa vérité de terrain de manière à réduire le temps passé à transcrire manuellement son corpus pour entraîner un premier modèle.

HTR-United : questions méthodologiques

15. Le projet HTR-United est né du constat qu'il faut mettre en commun la vérité de terrain pour permettre à chacun et chacune d'en bénéficier. Cela pose cependant de nombreuses questions méthodologiques que nous pouvons rappeler.

16. En premier lieu, le signalement, la documentation et les métadonnées. La description d'un jeu de données est cruciale pour permettre sa réutilisation par d'autres. En effet, on veut généralement savoir, par exemple, quelle est la langue utilisée dans les documents ou encore à quelle époque ils ont été rédigés. Ces informations permettent d'opérer un tri entre des lots de données pertinents pour composer une nouvelle vérité de terrain et ceux qui ne le sont pas. Au fil de nos réflexions sur l'écosystème HTR-United, nous avons élaboré un modèle de description[4] des jeux de données qui reprend notamment les champs suivants : la licence ; la langue ; le système d'écriture (ou alphabet) ; le nombre de mains[5] ou de polices et leur proportion ; la période couverte ; ou encore, l'importance matérielle (c'est-à-dire le volume). À ces éléments s'ajoutent les informations qui permettent d'identifier et de citer un jeu de données[6]. Le modèle de données fournit des indications sur la manière de remplir les champs correspondants en proposant, lorsque c'est possible, des listes fermées de valeurs. Cela permet de définir des pistes pour aboutir à une uniformisation des descriptions pour les cas complexes les plus courants. Par exemple, nous proposons une liste fermée pour décrire la manière dont les deux principaux états du texte sont représentés au sein d'un jeu de données (« *only-manuscript* », « *only-typed* », « *mainly-manuscript* », « *mainly-typed* » ou encore « *evenly-mixed* »), ou bien pour quantifier le nombre de mains représentées. HTR-United propose d'ajuster la quantification du nombre de mains à l'échelle des fichiers ou des dossiers avec des valeurs comme « *one-per-file* » (une main par fichier) ou « *one-per-folder* » (une main par dossier). Considérant d'une part que ce n'est pas le nombre exact de mains, mais l'importance de la variation des écritures qui importe, et que

4. HTR-United. (2021). HTR-United Schema (v. 15-10-2021). Alix Chagué & Thibault Clérice (éds.). URL : https://htr-united.github.io/schema/2021-10-15/schema.json (consulté le 10/02/2022).
5. Une « main » correspond à l'écriture d'un individu, donc à une variation d'écriture.
6. Nous proposons pour cela de fournir les informations (nom, prénom, rôle) permettant de citer l'ensemble des personnes ayant contribué à la création d'un jeu de vérité de terrain, notamment à travers les rôles « *transcriber* », « *aligner* », « *project-manager* » ou « *support* ».

d'autre part, il n'est parfois pas possible de quantifier avec précision cette variation, nous proposons trois options : « *one* », « *few* » (10 mains ou moins) ou « *many* » (plus de 10 mains). La précision de cette quantification est indiquée par un autre champ dont les valeurs peuvent être « *exact* » ou « *estimated* ».

17. Évoquons en deuxième lieu les standards qui sont un autre aspect méthodologique à prendre en compte. PAGE et ALTO constituent au moins deux exemples de standards, mais il faut noter qu'ils se déclinent chacun en plusieurs versions. Faut-il s'en tenir à un standard et une version uniques, et si oui, lesquels ? Est-il seulement possible de répondre à cette question alors que les logiciels continuent d'évoluer ? Par exemple, jusqu'à la publication de la version 1.5.0 de Transkribus en mars 2021, l'application de bureau (*desktop*) proposait d'exporter des données au format XML ALTO 2 et au format XML PAGE. Avec la version 1.5.0, le logiciel est soudainement passé à la version 4.2 d'ALTO, pour l'export et l'import des données, sans assurer de rétrocompatibilité avec ALTO 2[7]. On pourrait être tenté de penser que les modèles sont de ce point de vue plus robustes que les données, mais il faut noter qu'en janvier 2020, lorsque Kraken est passé à sa version 3, en permettant alors d'entraîner des modèles de segmentation en plus des modèles de transcription, tous les modèles produits avec les versions 2.x du logiciel ont cessé d'être compatibles avec les versions ultérieures.

18. Enfin, un troisième aspect méthodologique important : le contrôle de la qualité d'un jeu de données. Ce sont en général les objectifs du projet pour lequel un modèle est entraîné qui définissent la qualité attendue pour la vérité de terrain. Des invariants permettent toutefois d'établir plusieurs critères : l'homogénéité des règles suivies pour la production des données ; la fidélité de la transcription par rapport à l'image ; et sa capacité à s'adapter aux objectifs d'un autre projet. Dans un corpus de transcription comme celui créé à l'occasion de l'édition des journaux d'Eugène Wilhelm [Schlagdenhauffen, 2020] des passages rédigés en alphabet grec ont été transcrits en alphabet latin. Cela est justifié par le porteur de ce projet, mais constitue néanmoins une transcription qui diffère de ce que l'image originale contient : cela peut poser un problème pour une réutilisation dans le cadre d'un projet prévoyant de produire un modèle capable de distinguer alphabet grec et alphabet latin. Cette vérité de terrain potentielle est-elle pour autant de mauvaise qualité ? Non. En fait, ce qui importe, c'est qu'*a minima* l'information concernant la pratique de transcription suivie soit documentée afin qu'elle puisse être prise en compte par une personne ré-utilisant de telles données. Idéalement, ce genre de problématique est pris en charge dans le cadre d'un Plan de Gestion des Données (PGD) qui décrit le contexte de production des données et les règles de transcription établies avant la campagne de transcription, ou bien actualisées durant sa conduite.

7. Il est heureusement possible de mettre à niveau les jeux de données publiés avant mars 2021, à l'aide de script XSLT par exemple, sous réserve que toutes les informations attendues par ALTO 4 soient présentes.

Un projet adossé à GitHub

19. Le projet HTR-United a été mis en place sous la forme d'une organisation GitHub en octobre 2020. Il s'agit d'une entité propre à la plateforme permettant à plusieurs utilisateurs et utilisatrices de se rassembler autour d'un projet commun se déployant sur différents répertoires de travail. Le fait de s'appuyer sur une plateforme comme GitHub présente plusieurs avantages, dont la facilité d'y mettre en place un travail collaboratif dépassant les limites d'un projet donné ainsi que la possibilité de gérer finement plusieurs versions de travail et d'en faire coexister plusieurs simultanément.

20. L'organisation Github HTR-United est composée de plusieurs répertoires (figure 4) dont le principal, nommé « htr-united » [HTR-United, 2020/2022], contient un catalogue prenant la forme d'un fichier de texte YAML (htr-united.yml) dont le contenu est généré automatiquement grâce au moissonnage des métadonnées fournies par les contributeurs et contributrices. Les contributions sont ajoutées selon deux modalités : soit sous la forme d'un fichier de métadonnées créé, dans le répertoire principal, dans un dossier nommé « catalog », soit sous la forme d'un répertoire à part, créé au sein de l'organisation Github HTR-United.

21. Dans le dossier « catalog », un dossier est attribué à chaque projet contributeur, dans lequel on crée un fichier YAML par jeu de données. Par exemple, le projet e-ditiones a créé le corpus de vérité de terrain « OCR17+ » [Gabay *et al.*, 2020 ; Jahan et Gabay, 2021], il existe donc dans le dossier « catalog », un dossier nommé « e-ditiones » contenant un fichier de métadonnées nommé « ocr17plus.yml ».

22. Dans les répertoires satellites, plusieurs éléments sont attendus : des transcriptions alignées avec des images, enregistrées dans un format standard comme XML PAGE ou XML ALTO ; les images ou bien les informations permettant d'accéder aux images (par exemple par un lien vers un manifeste IIIF) ; un document de présentation du corpus et de son contexte de production sous la forme d'un fichier « readme.md », donnant autant d'informations que possible, y compris sur l'architecture du dossier de dépôt ; et enfin un fichier de métadonnées intitulé « htr-united.yml »[8]. Notons que dans le cadre de la publication d'un *data paper* décrivant un jeu de données pour l'entraînement de modèles HTR, notre structure propose ainsi un modèle opérationnel pouvant servir de base à la structuration des informations[9]. Nous définissons un ensemble d'éléments de documentation qui sont cruciaux pour comprendre et réutiliser ces données.

8. Nous proposons un modèle de répertoire de dépôt de vérité de terrain accessible via https://github.com/HTR-United/template-htr-united-datarepo (consulté le 16/03/2022).

9. On peut se référer à l'exemple du gabarit proposé par le *Journal of Digital Humanities* pour les *data papers* [Hengchen et Pedrazzini, 2022]. Il propose des items de descriptions similaires à ceux demandés par HTR-United.

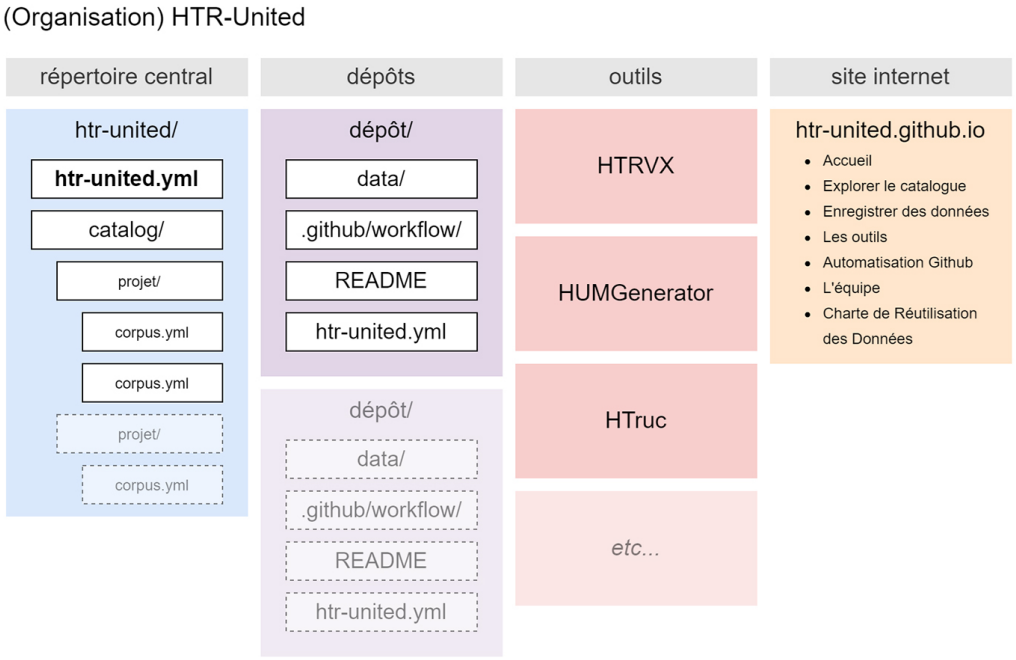


Figure 4. Schématisation des répertoires rassemblés dans l'organisation HTR-United et de leurs contenus

23. Les fichiers de métadonnées sont conformes au modèle de description évoqué plus haut. Un formulaire accessible depuis le site Internet du projet[10] aide à leur génération. Ce formulaire, ainsi que l'utilisation d'un format léger comme YAML, entend faciliter la création des descriptions par des personnes ne disposant pas des connaissances suffisantes pour produire des fichiers à structure plus complexe comme XML. YAML est en outre un format facile à analyser automatiquement : il permet de générer le catalogue principal recensant l'ensemble des dépôts, de contrôler la validité du contenu de certains champs, et surtout d'alimenter la page d'exploration du catalogue proposée sur le site Internet du projet[11].

24. Pour alimenter ces fichiers de métadonnées, il est possible d'automatiser le calcul des valeurs des champs liés à l'importance matérielle. C'est l'ambition de HTR-United Metadata Generator (HUM Generator) [Clérice et Chagué, 2021], un processus qui analyse les fichiers XML déposés afin de relever le nombre de pages, de lignes et de caractères constituant un lot de vérité de terrain. Combiner ces métriques est important, car, lorsqu'elles sont exprimées individuellement, elles ont peu de signification : le nombre de caractères dans une ligne et le nombre de lignes dans une page sont très variables en fonction des types de documents. Pour renseigner sur la taille réelle d'un lot de vérité de terrain, il faut donc les associer.

25. La transparence permise par GitHub signifie que, lorsqu'une personne publie sa vérité de terrain et la signale dans le catalogue HTR-United, sous réserve qu'elle soit libre de droits, une autre personne peut l'utiliser pour son projet et la citer, ou bien la mettre à jour ou la convertir pour la rendre compatible avec son logiciel

10. Le formulaire est accessible via l'URL suivante : https://htr-united.github.io/document-your-data.html.

11. Cette page est accessible via l'URL suivante : https://htr-united.github.io/catalog.html (consulté le 16/03/2022).

et la re-publier comme une version alternative du jeu de données initial[12]. À l'échelle d'un dépôt, cela permet aussi de mettre en place un mécanisme de publication progressive de la vérité de terrain : il n'est pas nécessaire d'attendre la forme la plus aboutie du jeu de données pour le publier, car on peut versionner le répertoire et mettre à jour progressivement le contenu des données ou bien sa documentation.

Un soutien au contrôle qualité

26. La production de vérité de terrain en HTR repose sur trois piliers : une production d'annotations – le texte, la segmentation –, sa formalisation – en XML, avec différents jeux de caractères –, et son intégration dans un réseau de production – à travers des ontologies de segmentation, des schémas et des choix d'encodage (figure 5). Si la première section relève principalement de données à vérification qualitative, l'ensemble des autres informations correspond à des éléments dont la validation peut être prise en charge par la machine. Dans ce cadre, afin d'assurer à la fois la qualité des données et réduire le temps passé à leur vérification formelle, HTR-United et le projet CREMMA travaillent à la mise à disposition de divers outils dits « d'intégration continue ».

27. L'intégration continue consiste au lancement automatisé et externalisé[13] de tests, voire de compilations[14] au moment de la synchronisation d'un dépôt tel que ceux de GitHub[15] : elle permet par son caractère décentralisé de produire une vérification publique de la qualité des données et du code à chaque modification. Cette pratique reste encore assez rare dans le domaine des données en humanités numériques, mais connaît une progression sur les dernières années [Almas et Clérice, 2017 ; Ferger et Hedeland, 2020].

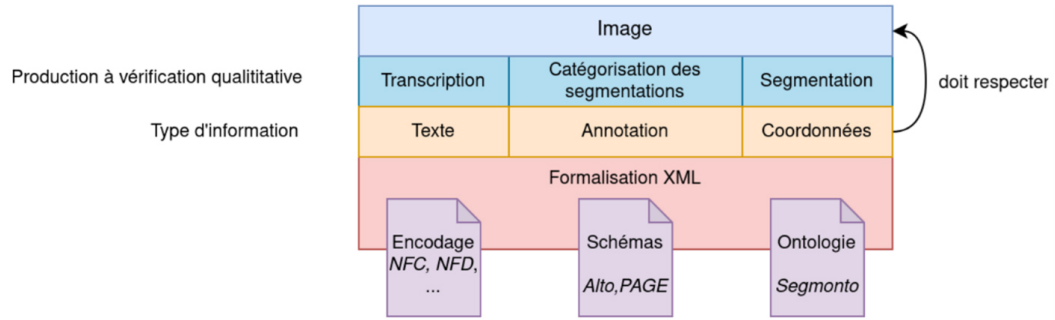


Figure 5. Typologie des informations et leurs relations dans le cadre de vérité de terrain

28. HTR-United propose l'utilisation de trois outils ayant chacun des objectifs partagés :

12. Pour comprendre le fonctionnement des forks dans GitHub : https://docs.github.com/en/get-started/quickstart/fork-a-repo [Github Inc., 2021].

13. Ces tests sont nécessairement lancés sur des machines vierges, permettant ainsi un test « objectif » des données : le même code est lancé indépendamment de toute particularité de paramétrage des ordinateurs de chacun et chacune.

14. Dans notre cas, la phase compilation peut prendre la forme d'une normalisation automatisée ou de versionnage automatique du corpus.

15. GitHub propose son propre service d'Intégration Continue, *GitHub Actions*, mais d'autres existent : TravisCI, CircleCI, etc.

- ChocoMufin [Clérice et Pinche, 2021b],
- HTRVX [Clérice et Pinche, 2021a]
- Et *HTR United Metadata Generator* (*HUM Generator*) présenté plus haut (figure 6).

29. ChocoMufin a été développé originellement dans le cadre du corpus CREMMA Médiéval [Pinche et Clérice, 2021] afin de traquer la variation dans l'encodage des caractères médiévaux. En effet, les manuscrits médiévaux présentent une très grande diversité d'abréviations utilisant différents signes additionnels, qu'ils soient « nouveaux » (7 [et], 9 [con]) ou non (macrons, barres obliques, etc.), de ligatures et de caractères (s connaît au moins trois variations principales : ß, ſ, ꝛ). Or, maintenir de la constance dans la transcription pour choisir le « bon » caractère peut s'avérer difficile. En outre, les pratiques diffèrent d'un projet à l'autre[16]. Afin de rendre ces pratiques « interopérables », ChocoMufin vérifie chacune des lignes transcrites en fonction d'une table de valeurs autorisées propre à chaque dépôt. Cette vérification s'accompagne d'une table des nouveaux caractères apparus, qui peuvent alors être inclus à la table des caractères validés ou au contraire corrigés. Par ailleurs, cette table des caractères contient aussi une valeur de remplacement, permettant aux utilisateurs et utilisatrices de proposer des « simplifications » de leurs transcriptions, afin d'uniformiser les pratiques entre dépôts et projets.

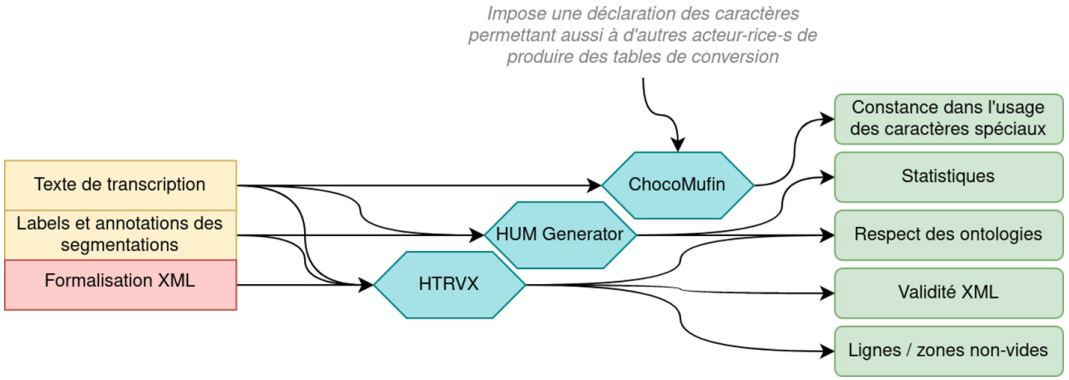


Figure 6. Informations, outils de contrôle et résultat de ces contrôles

30. Le deuxième outil est un simple outil de vérification des schémas, au format XSD. *HTRVX* s'appuie sur un schéma – nous fournissons uniquement un schéma pour l'ontologie SegmOnto [Gabay *et al.*, 2021] pour le moment – et permet alors la vérification de la validité du fichier en fonction des catégories de segmentation proposées par *SegmOnto*. Nos schémas incluent aussi une vérification d'absence de lignes vides, qui auraient pu échapper à l'œil des annotateurs et annotatrices, soit parce que la ligne était difficile à percevoir et à l'origine d'une erreur de segmentation, soit parce qu'elle a tout simplement été oubliée. Chaque fichier fait alors l'objet d'un rapport individualisé avec un regroupement lisible de l'ensemble des erreurs rencontrées.

16. Ce problème dépasse la période médiévale : d'une part, il est encore commun de trouver des abréviations à la période moderne dans les documents manuscrits, mais on trouve encore après cette période des variations graphiques tels S Long / S ou des ligatures (les éditions de la Pléiade présentent encore des st en ligature). Les abréviations type numéro, les guillemets, etc., peuvent aussi faire l'objet de variations d'un annotateur à une autre.

31. Ainsi, les personnes chargées de l'annotation et les gestionnaires de corpus de vérité de terrain réduisent le temps de maintenance et de recherche d'erreurs, en ne se concentrant que sur les logiques globales du corpus (normes de transcriptions, transcriptions) et en s'appuyant sur ces outils. Par ailleurs, les détails statistiques fournis par HUM Generator permettent de suivre la progression de la production de contenus, voire de fournir des badges publics informant les personnes découvrant les corpus de l'état de ceux-ci au moment de leur visite (figure 7).

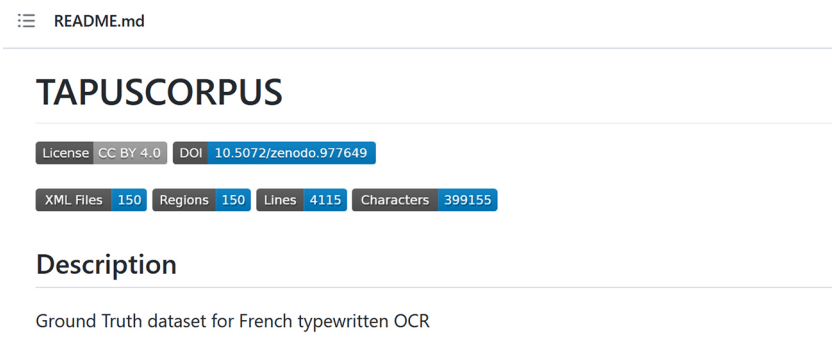


Figure 7. Badges statistiques automatiquement mis à jour pour un corpus HTR United

Conclusion

32. Mettre en commun la vérité de terrain est crucial pour permettre à la recherche d'avancer vers une meilleure intégration de la reconnaissance des écritures manuscrites dans les projets en humanités numériques. Une meilleure disponibilité des données passe par l'amélioration de la documentation sur leur contexte de production : on pense notamment aux règles de transcription, mais aussi aux objectifs du projet ainsi qu'aux éventuelles omissions volontaires. En plus de la documentation présente dans les dépôts de données d'entraînement, le format de publication des *data papers* constitue un excellent moyen de rendre publiques les stratégies d'échantillonnage, les pratiques de transcriptions et les éventuelles spécificités de ces jeux de données. Les efforts de structuration de cette documentation, opérés au fil de l'élaboration de l'environnement HTR-United, encouragent à l'établissement de pratiques homogènes pour la construction et la description de ces données. Ces échanges – publics et ouverts à toute contribution de la part de la communauté des utilisateurs et utilisatrices de ces technologies – pourraient finalement produire un modèle pour la rédaction de *data papers* spécifique aux données d'entraînement pour l'HTR.

33. Nous avons montré l'importance de mettre en place un recensement correct des jeux de données, passant notamment par l'établissement d'un modèle de métadonnées et d'outils facilitant son appropriation par des chercheurs et chercheuses aux profils variés. L'uniformisation des descriptions de la vérité de terrain permet d'interroger la communauté sur les critères permettant de filtrer efficacement plusieurs ensembles de données. Des initiatives similaires existent à d'autres niveaux : on peut mentionner à nouveau le projet SegmOnto, dont

l'ambition est de produire des modèles de segmentation et d'analyse de mise en page basés sur des données très diversifiées, mais suivant les mêmes règles d'annotation sémantique. L'objectif d'un tel projet est double : produire des modèles prêts à l'emploi et adaptés aux documents patrimoniaux manuscrits et imprimés, et partager des données permettant d'aboutir à ces modèles.

34. L'initiative que nous avons décrite s'appuie très fortement sur des mécanismes d'intégration continue et de versionnage qu'une plateforme telle que GitHub rend particulièrement propices. En plus d'inciter à la transparence des processus de constitution de la vérité de terrain, le paradigme adopté permet d'alléger la tâche de contrôle qualité en fixant des critères qu'il est possible d'adapter (absence ou tolérance des lignes ou des zones vides, conformité ou non au modèle sémantique SegmOnto, etc.) et en automatisant ces contrôles. S'appuyer sur une infrastructure privée n'est pas sans poser question, mais il existe fort heureusement des mécanismes d'archivage pérennes pour les répertoires GitHub, comme la Software Foundation, qui permettent à HTR-United d'offrir aux projets contributeurs des moyens de répondre aux objectifs d'accessibilité des principes FAIR. D'une manière générale, HTR-United contribue à la découvrabilité des jeux de données d'entraînement pour la transcription automatique, incite à l'emploi de standards garantissant l'interopérabilité et met en place les conditions de leur réutilisation par le biais du modèle de description implémenté. À mesure que la quantité de données signalées dans le catalogue grossit, on peut envisager que des institutions patrimoniales s'empareront de la question du recensement et de collecte de la vérité de terrain afin d'en pérenniser l'enregistrement.

35. Il est temps d'encourager la publication de ces données pour ce qu'elles sont : des données d'entraînement et pas seulement des transcriptions. Cela peut passer par la mise en place de chartes incitant au dépôt de la vérité de terrain en contrepartie de l'utilisation de ressources librement mises à disposition, comme ce sera par exemple le cas du serveur CREMMA, financé par le DIM MAP [DIM MAP, 2021]. Puisque nous visons de garantir une simplicité d'utilisation, HTR-United peut également être intégré dans les cursus universitaires qui forment à la transcription ou aux outils de versionnage. En effet, en réalisant une simple tâche d'alignement entre transcription et image, ou en mettant à jour un corpus, n'importe qui peut contribuer à cette initiative.

Nous proposons un modèle de répertoire de dépôt de vérité de terrain accessible via https://github.com/HTR-United/template-htr-united-datarepo

De l'entrepôt de données aux *data papers*. Retour sur l'expérience de Data Sciences Po

Alina Danciu, Anna Egea, Guillaume Garcia et Cyril Heude

1. Ce chapitre restitue la manière dont la mise en œuvre d'un entrepôt Dataverse mutualisé – data.sciencespo.fr – nous a poussés à investir la question des *data papers*. Ce compte rendu croise les expériences de différents acteurs (laboratoire, service transverse, centre de données) concernés par cette démarche. Nous montrerons comment notre politique en faveur des dépôts de jeux de données nous a confrontés à des obstacles liés au manque de moyens humains, et comment cela nous a conduits à nous intéresser aux *data papers* comme un outil stratégique pour renforcer l'accompagnement dans le processus de dépôt des données.

2. Nous remettrons d'abord en contexte les dispositifs qui ont précédé la mise en place de Data Sciences Po. Nous concentrerons notre propos sur cet entrepôt qui s'inscrit pleinement dans la politique de science ouverte de Sciences Po et abrite deux collections différentes, l'une destinée aux chercheurs et chercheuses de Sciences Po (collection Sciences Po, en auto-dépôt accompagné), et l'autre destinée à la communauté nationale et internationale de recherche en sciences sociales (banque de données du Centre de données socio-politiques [CDSP][1]). Nous partagerons notre retour d'expérience sur la mise en place de cet entrepôt, en mettant en lumière les coûts significatifs, principalement humains, indispensables à sa maintenance et à son animation. Rappelons que cet entrepôt a été mis en place en utilisant des ressources déjà existantes à Sciences Po et n'a pas bénéficié d'un budget dédié, ce qui aurait été souhaitable.

3. Nous soulignerons ensuite les problèmes que nous rencontrons plus particulièrement pour favoriser la documentation des données, quelles que soient leurs modalités de dépôt. L'entrepôt propose un service d'auto-dépôt accompagné et un service de curation des données qui est assuré pour le compte des déposants. Dans les deux cas, le renseignement des métadonnées est pris en charge quand cela est possible. Les équipes comptent sur la contribution, variable en pratique, des chercheurs et chercheuses pour apporter une contribution à ce travail. L'anonymisation des données est assurée par les ingénieurs du CDSP pour les jeux de données déposés au sein du laboratoire, et ce sont les déposants eux-mêmes qui anonymisent les références publiées dans la collection « auto-dépôt accompagné ».

4. Pour favoriser la réutilisation et la visibilité des données, nous nous dirigeons vers la mise en place de dispositifs

1. https://cdsp.sciences-po.fr/fr/

d'accompagnement à la publication de *data papers* ; inciter les chercheuses et chercheurs à auto-documenter leurs jeux de données constituant également un moyen d'optimiser le service d'accompagnement au dépôt sous contrainte de ressources limitées. Nous présenterons les solutions que nous commençons à mettre en place dans ce sens.

Quels sont les enjeux de la documentation des données ?

5. Depuis 2006, le dépôt et la diffusion des données en sciences sociales à Sciences Po étaient assurés par le CDSP (Unité d'appui à la recherche [UAR] du CNRS et de Sciences Po) qui a la particularité d'accueillir essentiellement des ingénieurs issus de différents métiers. La démarche a débuté avec des données d'enquêtes recueillies par des méthodes quantitatives – le plus souvent obtenues par la passation de questionnaires sur des échantillons représentatifs de la population générale ou de certains groupes sociaux spécifiques. L'outil qui permettait d'explorer en ligne ces données, NESSTAR[2], a été utilisé jusqu'en 2020, son évolution et sa maintenance n'étant plus assurées par le Norwegian Centre for Research Data. Un peu plus tard, le CDSP a conçu et construit la base de questions Quetelet, impliquant son utilisation en partenariat avec le réseau homonyme connu aujourd'hui sous le nom de Quetelet-Progedo-Diffusion[3]. L'outil permettait de faire des recherches dans le texte des questions, les codes et étiquettes des modalités de réponse, les noms et les étiquettes de variable. À partir de 2013, le CDSP a commencé à diffuser des données d'enquêtes qualitatives – obtenues par entretiens approfondis ou par observations[4]. L'outil construit et utilisé pour explorer les données en ligne s'appelle beQuali[5].

6. Cet ensemble de services s'adresse aussi bien aux chercheurs et équipes de recherche situés à Sciences Po qu'à l'extérieur ; le CDSP prend en charge en interne la curation des données[6] pour le compte des déposants, et la diffusion des données s'est effectuée depuis de nombreuses années depuis le portail Quetelet-Progedo-Diffusion. Elle se fait aujourd'hui depuis la nouvelle plateforme Data Sciences Po[7].

7. Data Sciences Po est né d'un projet aujourd'hui clos, Archipolis. Archipolis était un réseau de laboratoires de Sciences Po et d'autres universités en France (Lille, Grenoble, Lyon, Bordeaux) organisé en consortium d'Huma-Num[8] entre 2012 et 2016. Le projet visait à

2. http://www.nesstar.com

3. https://www.progedo.fr/donnees/quetelet-progedo-diffusion/
4. Une vingtaine d'enquêtes qualitatives ont été collectées.
5. https://bequali.fr. Les dépôts sont en cours de migration au sein d'une collection dédiée de la Banque de données sociopolitiques.
6. Par curation, on entend des opérations comme l'anonymisation des données, l'ajout de métadonnées, la migration des fichiers des données en formats libres.
7. https://data.sciencespo.fr/
8. https://www.huma-num.fr/les-consortiums-hn/

créer une dynamique de préservation et de valorisation d'archives d'enquêtes de terrain. Seules étaient concernées les archives d'enquêtes qualitatives – principalement par entretiens – situées dans un domaine disciplinaire hybride, les sciences sociales du politique – essentiellement en science politique et en sociologie (Duchesne *et al.*, 2014). Archipolis a notamment permis de réaliser des inventaires d'enquêtes dans les différents laboratoires membres[9] et de développer, en 2015, une collection dédiée sous le logiciel libre Dataverse. L'objectif était de sensibiliser les chercheurs et ingénieurs des laboratoires à l'ouverture des données de la recherche, en proposant une première étape consistant à inventorier et à décrire dans un entrepôt, les jeux de données qui y étaient produits. La diffusion des données elles-mêmes était prévue dans un second temps de développement du projet, mais l'arrêt du réseau en 2016, à la fin du financement du consortium, n'a pas permis de franchir cette étape.

8. Ces notices étaient organisées comme un ensemble de métadonnées relevant du standard international *Data Documentation Initiative* (DDI)[10].

Figure 1. Catalogue Archipolis

9. https://catalogues.cdsp.sciences-po.fr/dataverse/archipolis/. 238 notices ont été élaborées, sur les 8 laboratoires membres du réseau en 2016.

10. Le CDSP est fortement impliqué dans la communauté en charge du maintien et de la formation à ce standard de métadonnées en SHS. Il est par ailleurs membre de la DDI Alliance, organisme qui fait évoluer ce standard.

Figure 2. Exemple d'une notice d'enquête Archipolis

9. Les notices étaient renseignées par les ingénieurs des laboratoires du réseau ou des contractuels recrutés sur le projet, en lien avec les producteurs des enquêtes. Avec Archipolis nous installions l'idée d'un circuit reliant projets de recherche, publications et données, chercheurs et ingénieurs, et commencions à questionner la relation strictement personnelle entre les données et leurs producteurs qui prévalait jusqu'alors.

Le passage à l'entrepôt de données de la recherche Data Sciences Po

10. À Sciences Po, la politique de la science ouverte est portée par la Direction des ressources et de l'information scientifique (DRIS) conjointement avec la Direction scientifique (DS) et son Centre de données socio-politiques. Cette politique a abouti, entre autres, à la mise en place d'une archive ouverte institutionnelle, à la rédaction d'un texte-cadre et à une expertise développée autour de la préservation et la diffusion de données, à la mise en place d'entrepôts de données, à la participation à des réseaux nationaux et internationaux de gestion de la donnée, à la mise en place de plans de gestion de données, et à la rédaction de guides[11]. Par ailleurs, la DRIS anime un réseau interne de partage d'expériences ; le CDSP coordonne un groupe de travail international sur le standard de données Data Documentation Initiative et participe au comité CoreTrustSeal, chargé de l'attribution des certifications avec le même nom.

11. Vers 2016, l'idée a émergé de doter Sciences Po d'un entrepôt institutionnel de données de recherche. De nombreux acteurs ont été mobilisés pour développer Data Sciences Po et plusieurs univers professionnels sont articulés pour faire vivre le projet : des data managers, des data librarians, des juristes, des administrateurs système, des développeurs, ainsi que les laboratoires de recherche, qu'il s'agisse des chercheurs ou des personnels

11. https://sciencespo.libguides.com/donnees-de-la-recherche/

IST. Le CDSP a coordonné sa mise en place, en se basant sur l'expérience accumulée avec Dataverse dans le cadre du projet *Archipolis*. La mise en place de cet entrepôt, ainsi que sa maintenance et son animation exigent de multiples compétences et moyens humains non négligeables pour les institutions qui souhaitent se lancer dans cette aventure (traitement de métadonnées et données, expertises juridiques, développement et maintenance d'un entrepôt, accompagnement des déposants).

12. Deux modalités de dépôt et de curation des données sont désormais disponibles.

Figure 3. Page d'accueil de *Data Sciences Po*

13. Une première collection[12] accueille les données produites par les chercheurs et chercheuses affiliés à Sciences Po (données d'enquête, bases de données, etc.), sur le principe d'auto-dépôt accompagné par la DRIS, en collaboration avec les personnels des laboratoires[13].

12. https://data.sciencespo.fr/dataverse/adscpo/

13. On y retrouve des jeux de données moissonnés depuis le Dataverse de Harvard (évolution d'affiliation de la chercheuse concernée).

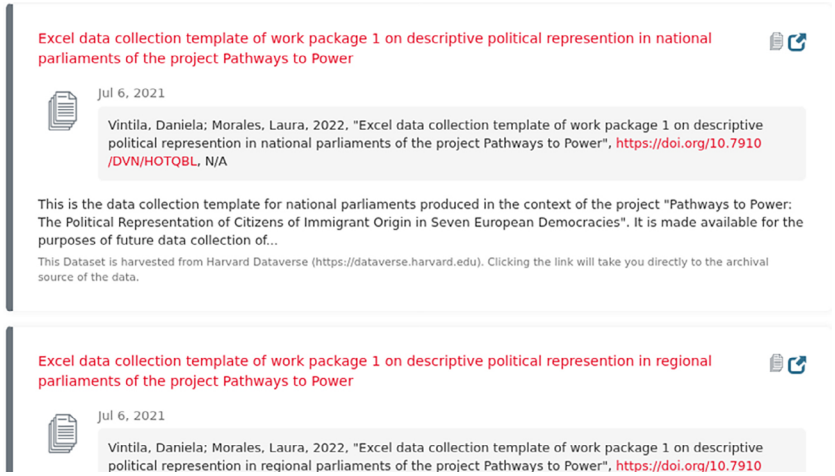


Figure 4. Catalogue de la collection « dépôt accompagné » de Data Sciences Po

14. La Banque de données du CDSP comprend le catalogue d'enquêtes en sciences humaines et sociales traitées et contextualisées par les ingénieurs du CDSP depuis 2006[14]. Ont accès à ce service des producteurs de données localisés en France et à l'international[15].

14. https://data.sciencespo.fr/dataverse/cdsp/. Le catalogue est composé de 400 enquêtes et bases de données.

15. La Banque de données du CDSP a obtenu la certification CoreTrustSeal en 2024 ; il s'agit actuellement du seul entrepôt spécialisé en enquêtes SHS provenant de méthodes quantitatives et qualitatives en France à avoir obtenu cette certification. Par ailleurs, la banque de données est l'un des entrepôts de confiance mentionnés dans la note établie par le Collège des données de la recherche du Comité pour la science ouverte (Référence : https://www.ouvrirlascience.fr/donnees-de-la-recherche-comment-choisir-un-entrepot-de-confiance/).

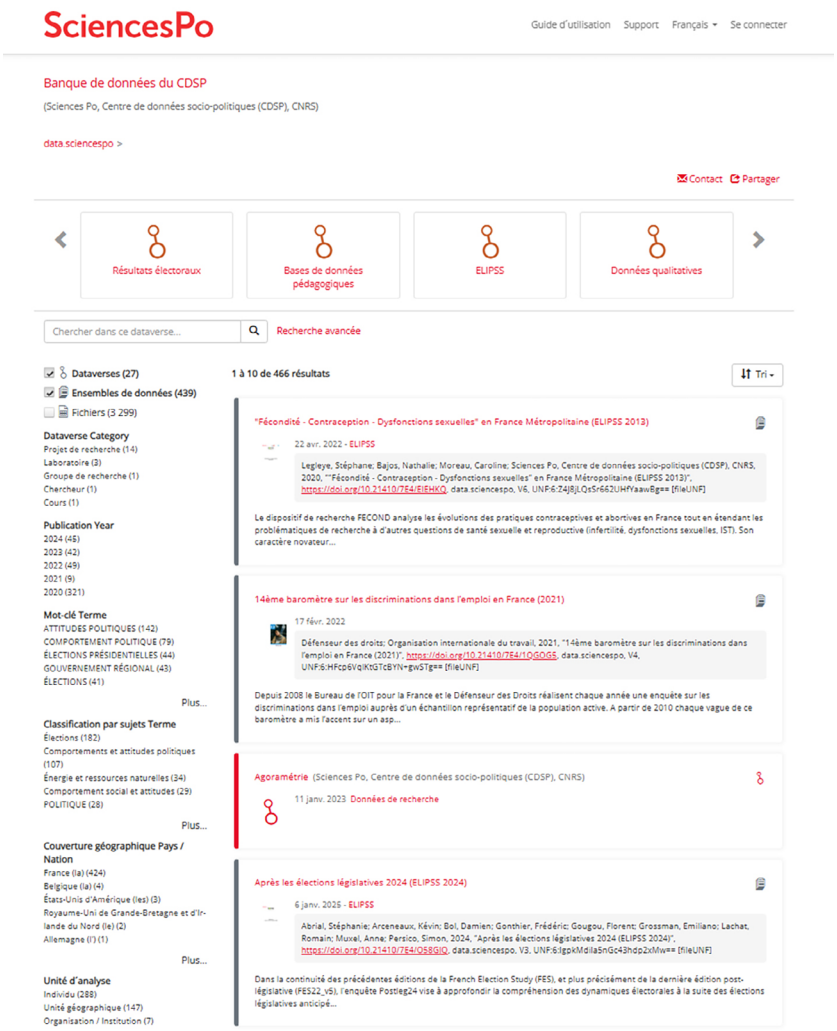


Figure 5. Catalogue de la collection « banque de données du CDSP » de Data Sciences Po

15. Quelles que soient les modalités de dépôt, les données sont contextualisées en respectant les normes internationales en vigueur, notamment le standard DDI. Un DOI est assigné à chaque jeu de données, favorisant ainsi la citabilité et le crédit académique.

16. Les avantages de cet entrepôt qui respecte les principes FAIR sont multiples. Il est moissonnable par d'autres entrepôts et référencé dans les moteurs de recherche, ce qui permet d'accroître significativement la visibilité des données pour les réutilisateurs[16]. Par ailleurs, l'entrepôt répond aux besoins de la communauté des producteurs et utilisateurs de données de recherche au regard des exigences en matière de plans de gestion de données (projets ANR, H2020, etc.). L'entrepôt accepte tous types de données et de formats[17]. Il a été développé sur la base du même logiciel que l'entrepôt national de la recherche, Recherche Data Gouv[18], ce qui facilite le moissonnage par ce dernier.

17. L'outil technique comme les dispositifs de sensibilisation pour (auto)déposer les données existent, et des expertises comme des ressources peuvent être mobilisées pour appuyer différents aspects de la curation des données ; néanmoins la documentation reste une activité complexe et coûteuse en temps et ressources humaines.

16. Data Sciences Po est référencé par exemple par le registre des données de la recherche re3Data.

17. Outre l'intégration de la plupart des fonctionnalités génériques de Dataverse destinées à valoriser la visibilité, la citation et la réutilisation des jeux de données, des collections par projet/équipe de recherche sont possibles, tout comme les différents types d'accès (ouvert, sur demande, restreint). Sur les fonctionnalités de Dataverse, voir Szabo 2019.

18. https://www.ouvrirlascience.fr/recherche-data-gouv-plateforme-nationale-federee-des-donnees-de-la-recherche/

Documenter des jeux de données, une activité complexe et coûteuse

18. Nous avons d'une part un centre de données qui assure la documentation et la diffusion des données, mais dont les services sont disponibles sous réserve d'acceptation des projets de dépôt, faute de moyens suffisants pour tout prendre en charge. D'autre part, nous proposons également un service sans sélection préalable, offrant aux déposants le choix entre une prise en charge complète du dépôt par les services de soutien ou une formation à l'auto-dépôt, leur permettant ainsi de s'exercer et de gagner en autonomie. Notre expérience donne à voir les conséquences de ces deux situations complémentaires en matière de prise en charge de la documentation des données.

Gérer la documentation pour le compte des déposants

19. Gérer ce service pour le compte des déposants permet d'aboutir à une documentation enrichie des jeux de données, mais a un coût non négligeable.

20. S'agissant des jeux de données dont la curation est d'emblée prise en charge par le CDSP[19], la situation est relativement confortable pour les producteurs déposants. L'équipe du CDSP, dont c'est le cœur de métier, dispose de ressources spécialisées, tant en termes d'effectifs que d'expertises, couvrant la production, le traitement et l'anonymisation des données quantitatives et qualitatives, ainsi que la conception et la gestion d'entrepôts de données et de bases de questions. Les déposants, tout en étant associés à chaque étape, n'ont pas à réaliser directement ce travail. Avant toute chose, les ingénieurs du CDSP s'assurent que la confidentialité des « répondants » ou des « enquêtés » est respectée. Ensuite, les données sont documentées de manière fine à l'aide du standard de métadonnées DDI. Le CDSP utilise ici le modèle de métadonnées du CESSDA[20] (réseau des centres de données européens en données SHS), grâce auquel il est possible de renseigner des informations au niveau de l'enquête elle-même (et donc accéder à un certain niveau de connaissance du protocole de recherche) mais aussi documenter plus finement les données, en particulier les données « quantitatives ». Ces dernières sont décrites à l'aide de logiciels tels que NESSTAR, Colectica ou R, et cette documentation approfondie facilite à la fois la recherche parmi les questions et variables ainsi que la réutilisation des données. Par ailleurs, le CDSP utilise des vocabulaires contrôlés comme ELSST[21] et ceux de la DDI Alliance, les seconds étant traduits et maintenus en français par le laboratoire. Les données « qualitatives » sont, elles, documentées avec des outils *in-house*.

19. Les données mises à disposition par le CDSP servent régulièrement dans des projets d'analyse secondaire qui donnent lieu à des publications ou qui sont utilisés dans des cours d'enseignement des méthodes quantitatives ou qualitatives. Un exemple récent d'ouvrage est *Controlling the Electoral Marketplace* (van Spanje 2018), qui utilise des données à thématique électorale diffusées par le CDSP.

20. https://vocabularies.cessda.eu/

21. https://elsst.cessda.eu/

Figure 6. Exemple d'une notice d'enquête de la *Banque de données sociopolitiques*

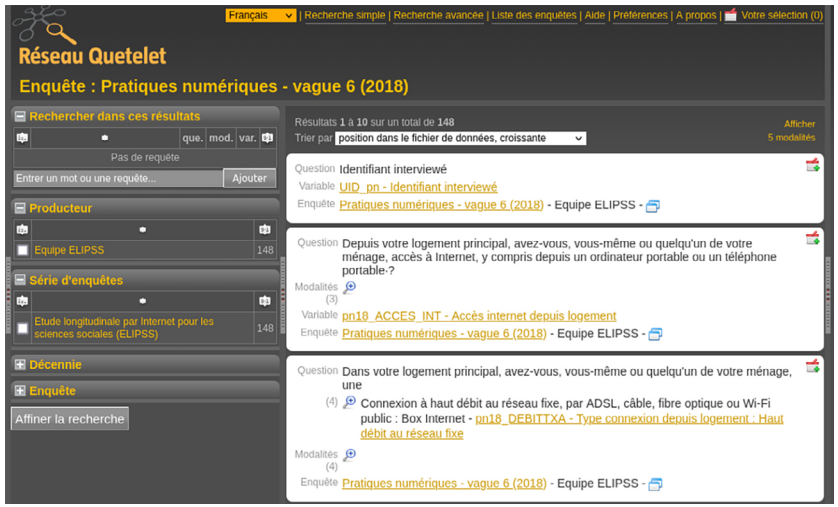


Figure 7. Exemple d'une documentation très fine au niveau des variables quantitatives.
Crédit : base de questions *Quetelet*[22]

21. Pour les enquêtes qualitatives, cette tâche implique la rédaction d'une « enquête sur l'enquête », réalisée en collaboration avec les chercheurs déposants. Ce document,

22. http://bdq.quetelet.progedo.fr/fr/Questions_et_variables_d_une_enquete/FR/1998/. Cette plateforme n'est plus d'actualité dans cette forme. Le CDSP travaille à la mise en place d'une nouvelle base de questions pour ses enquêtes et a récemment obtenu un financement dans le cadre de l'appel Réso 2024 à ce sujet (https://anr.fr/fr/detail/call/recherche-sur-les-pratiques-et-enjeux-de-la-science-ouverte-reso/).

Files Description

Dataset contains 3 file(s)

attributs	
# Cases	71
# Variable(s)	33

File Content
3 fichiers principaux sont diffusés dans le cadre de cette enquête.

Le fichier "attributs" est le premier de cette série. Il représente, par avocat, l'ensemble des attributs et des variables strictement individuelles. On retrouve ainsi dans ce fichier les propriétés socio-démographiques, les choix normatifs concernant l'avenir de la firme, ou encore les données comptables sur les performances économiques.

Ce fichier "attributs" peut être complété par le dossier comportant les matrices "conseil", "amitié" et "collègues", nommé "lawyer_networks".
Dans ce dossier, chacune des matrices, sous forme de fichier csv, présente en lignes et en colonnes l'ensemble des avocats. A l'intersection d'une ligne et d'une colonne est alors inscrit "0" ou "1" suivant qu'une relation existe ou non entre les deux avocats, si cette personne l'a déclaré dans son questionnaire.

En plus de ces trois matrices qui concernent l'ensemble du cabinet, il existe aussi une matrice "influence", qui correspond pour chacun des 36 associés aux relations d'écoute au sein de l'Assemblée générale des associés (qui écoute qui).

triplets	
# Cases	3043
# Variable(s)	46

File Content
Utile pour comprendre les mécanismes de contrôle latéral, le fichier "triplets" représente les relations déclarées entre les avocats.
Ces relations sont de type : "pour faire pression sur J, moi I, j'utiliserais K".
Les informations présentes dans ce fichier permettent de mieux comprendre les relations au sein du cabinet.

La question posée était la suivante (Lazega et Lebeaux 1995) :
Voici la liste de tous les associés de votre cabinet. Imaginez que vous êtes le directeur (managing partner). Vous vous rendez compte que l'un de vos associés a des problèmes personnels qui ont des répercussions négatives sur sa productivité. Ces problèmes peuvent être de toutes sortes : alcoolisme, dépression, divorce, etc. En tant que directeur, c'est à vous de vous préoccuper de cette situation. Vous cherchez parmi les associés de cette personne en difficulté un ou des collègues qui iraient lui parler discrètement et confidentiellement pour savoir ce qui se passe, et pour voir ce que le cabinet peut faire pour aider et limiter les dégâts. Vous ne voulez pas le faire vous-même parce qu 'il faut que la démarche reste informelle, et votre statut de directeur pourrait être gênant à cet égard. Ma question est la suivante : à qui, parmi tous les autres associés, demande riez-vous d'aller parler à Associé No 1, si c'est lui qui est en difficulté ?

Les informations du jeu de données "triplets" sont aussi présentes au sein du dossier "pression_laterale".
Ce dossier contient 36 matrices (une par associé) de 36*36 cases (leviers * cibles). Chaque matrice représente les données brutes sur les choix de leviers (k) effectués par chaque répondant (i) pour faire pression sur chaque cible (j). L'information contenue dans ces matrices est aussi représentée dans le fichier "triplets" où chaque observation équivaut à un choix de levier (n=3043). Dans la littérature ces bases de données s'appellent « three-way datasets ». Elles sont ici utilisées pour reconstituer et analyser le régime de contrôle latéral entre pairs. Pour cela, ces données sont enrichies par les attributs ou caractéristiques personnelles des acteurs et par la combinatoire des choix de leviers et de l'existence (ou non) de relation de collaboration, de conseil et/ou d'amitié entre répondant et levier, répondant et cible ainsi que levier et cible.

Figure 8. Exemple de documentation
des variables de l'enquête « Le phénomène collégial »[23]

Sommaire

Figure 9. Sommaire d'un rapport
portant sur l'enquête « Choisir son école »[24]

23. https://cdsp.sciences-po.fr/fr/ressources-en-ligne/ressource/fr.cdsp.ddi.phenome-necollegial1991/

24. https://bequali.fr/fr/les-enquetes/lenquete-sur-lenquete/cdsp_bequali_s1/

souvent un rapport de plusieurs dizaines de pages, vise à recontextualiser les aspects du processus de recherche qui demeurent obscurs ou insuffisamment explicités par les archives elles-mêmes ou par les publications issues de l'enquête. Le document retrace la genèse de l'enquête, ses ancrages théoriques, la réalisation du terrain, ou encore l'analyse des données.

Retour d'expérience du dépôt accompagné opar les *data librarians* et les laboratoires

22. Une tournée de promotion de Data Sciences Po dans tous les laboratoires de Sciences Po au premier semestre 2021 a permis de promouvoir une nouvelle modalité de documentation des données et de travail collaboratif : le dépôt accompagné, tout en la conjuguant à l'offre de services déjà existante. L'organisation de cette tournée a profité du réseau initié par le groupe inter-labos et services transverses animé par la DRIS[25]. Les interventions ont eu lieu en français ou en anglais dans divers cadres : assemblées générales de laboratoires (modalité efficace, car elle permet de s'insérer dans un ordre du jour plus large et d'atteindre également les personnes *a priori* moins intéressées par le sujet), séminaires de recherche, réunions du personnel, réunions *ad hoc* (modalité moins efficace, car elle n'attire que les personnes déjà convaincues). Des guides d'aide au dépôt et à la recherche de données dans l'entrepôt ont été mis à jour et traduits en anglais pour l'occasion[26].

23. Cette tournée a permis d'identifier des chercheurs pionniers. Certains argumentant en faveur de l'ouverture des données collectées auprès de leurs collègues pendant les démonstrations de l'outil, et ce, dans des disciplines *a priori* peu favorables à l'open science pour des raisons économiques (Droit). D'autres souhaitant faire apparaître dans l'entrepôt leurs nombreux jeux de données visibles dans d'autres entrepôts internationaux afin de les utiliser comme levier pour susciter l'intérêt et l'engagement de leurs collègues. D'autres, quant à eux, ont déjà commencé à déposer leurs propres données. Les dépôts effectués par des chercheurs[27] en amont de l'intervention ont conduit les collègues du même laboratoire à faire de même dans les mois qui ont suivi. Avant 2021, treize chercheurs de Sciences Po se sont lancés dans l'aventure, rejoints par trois chercheurs extérieurs, témoignant ainsi de l'importance des collaborations. Les premiers dépôts ont permis de constater que, parmi la centaine de fichiers disponibles, 84 % sont des données, et 16 % des éléments de contextualisation (fichiers « lisez-moi », par exemple). Les dépôts suivent les normes DDI, ELSST, Loterre, Unesco Thesaurus[28]. Les données quantitatives

25. https://sciencespo.libguides.com/donnees-de-la-recherche/encontacts/

26. Ont ainsi été publiés deux guides : un guide pour déposer ses données : https://data.sciencespo.fr/misc/guides/Guide_DEPOSIT_DSCPO_20200312.pdf ainsi qu'un guide pour trouver des données : https://data.sciencespo.fr/misc/guides/Guide_DOWNLOAD_DSCPO_20200312.pdf.

27. Cf. le dépôt réalisé par Hélène Le Bail (CERI), décrit ci-dessous.

28. Voir par exemple Barone (2021).

sont majoritaires en nombre de jeux de données, et les données qualitatives majoritaires en nombre de fichiers. Le lien dynamique entre les données d'appui et les publications est rendu possible, sur HAL, par un champ dédié (« Données associées - Ajoutez les identifiants DOI fournis par l'entrepôt où vos données sont archivées »). Des jeux de données s'enrichissent au fil des mois et des vagues d'entretien (deux fichiers en mars 2021, 23 fichiers en février 2022 par exemple) et sont très téléchargés : 1 600 téléchargements[29].

24. Des collections sont dédiées à des projets de recherche : par exemple l'*Enquête inter-associative sur l'impact de la Loi prostitution* (Le Bail et Giametta, 2021), menée par Hélène Le Bail, chargée de recherche CNRS au CERI (UMR Sciences Po/CNRS). L'objectif de cette enquête est de documenter l'impact de la loi de 2016 « visant à renforcer la lutte contre le système prostitutionnel et à accompagner les personnes prostituées » sur leurs conditions de vie et de travail. Au sein de l'entrepôt, l'enquête fait l'objet d'une enveloppe dédiée de trois jeux de données : 70 entretiens et témoignages courts de travailleurs et travailleuses du sexe en cinq langues ; 25 entretiens et *focus groups* avec des associations de terrain ; documentation de l'enquête (protocole d'enquête, grille d'enquête, grille de questionnaire, méthodologie de sondage, charte d'utilisation). Tous les niveaux d'accès (ouvert, sur demande, restreint) sont ici représentés.

> Hélène Le Bail à propos de sa démarche de dépôt : « Je trouve cela intéressant de rendre accessibles à tout le monde les documents de mise en place de l'enquête. Cela peut permettre, d'une part, à des personnes de s'inspirer du protocole d'enquête, voire de reproduire certaines questions dans le cadre d'enquêtes similaires. Sur un sujet polémique comme celui de la législation sur le travail du sexe et sur un terrain difficile d'accès, il me semble important de rendre visible comment on peut arriver à mettre en place une enquête de terrain. Par ailleurs, mettre les données dans *Data Sciences Po* permet de porter à la connaissance d'équipes de recherche sur le sujet l'existence de ces données et leur composition et de leur proposer d'entrer en contact avec ceux qui ont produit ces données pour d'éventuels accords de revalorisation »[30].
>
> Plein feu sur une chercheuse ayant pratiqué le dépôt accompagné

25. Par ailleurs, une réflexion pour inclure les plans de gestion de données (PGD) dans l'entrepôt est en cours : ils pourraient servir de fichiers de contextualisation à d'éventuels dépôts, voire nourrir le contenu de futurs *data papers* et faire gagner du temps aux chercheurs et aux chercheuses. L'accompagnement à la rédaction de PGD constitue un moment fort de sensibilisation au dépôt et aux *data papers* qui intègrent à ce titre la section « Partage » ou « Accès » du PGD. Des chercheurs de Sciences Po se sont d'ores et déjà engagés à écrire des *data papers* dès le début de leur projet. Par ailleurs, un réseau de correspondants *Data Sciences Po* au sein des laboratoires a été créé et formé. Ce réseau s'appuie sur un groupe inter-laboratoires et services transverses, animé par la DRIS, qui réunit des expertises variées (archiviste, documentaliste, statisticien, développeur, webmaster,

29. Voir par exemple Brouard *et al.* (2020).

30. Rencontre avec la chercheuse en septembre 2021.

cartographe, secrétaire générale). L'objectif de ce groupe est de promouvoir une approche collaborative pour résoudre collectivement les problèmes rencontrés sur le terrain.

26. L'intégration des données repose sur une collaboration entre le « data librarian » et les personnels des laboratoires. Ce travail englobe plusieurs étapes : analyse des données et réflexion sur leur présentation dans l'interface pour garantir un dépôt clair et accessible aux utilisateurs extérieurs, sélection des données, structuration hiérarchique des dossiers, nommage cohérent des fichiers, conversion des formats si nécessaire, définition de métadonnées pertinentes, anonymisation éventuelle des données, conseils sur les licences de diffusion, ainsi que la collecte des informations pertinentes issues des plans de gestion des données (PGD).

27. L'encouragement à l'auto-dépôt comme les services de documentation et de curation des données visent des objectifs communs autour du partage des données. Le partage des données représente un travail important, qui se surajoute aux charges de travail habituelles qui incombent aux chercheurs ou aux ingénieurs et bibliothécaires. Le travail de documentation, tel qu'il a été organisé jusqu'ici, devient aujourd'hui difficile à assumer dans une perspective de partage d'un nombre croissant de jeux de données. Ce travail de l'ombre est, de plus, peu valorisé, au mieux par la diffusion, sous le statut de littérature grise, dans HAL ou dans Data Sciences Po.

Faire des *data papers* une opportunité pour documenter les jeux de données, un véritable défi

Favoriser la pratique de rédaction des *data papers*

28. En mettant de côté la question de la mobilisation des chercheurs et en se focalisant sur celles et ceux déjà sensibilisés au partage des données, un enjeu majeur demeure : la documentation des données. Comment rendre cet exercice à la fois attractif et efficace, tant pour les déposants que pour notre équipe ? Une piste explorée est le développement de la pratique de rédaction de *data papers*. Selon le site CoopIST, le *data paper* est un format d'article qui « informe la communauté scientifique de la disponibilité de jeux de données et de leur potentiel pour des utilisations futures. Contrairement à un article de recherche classique, le *data paper* décrit uniquement des données scientifiques et les circonstances et méthodes de leur collecte »[31]. À la différence des situations où les chercheurs sont sollicités pour fournir des informations destinées à compléter des notices DDI ou à rédiger des documents de contextualisation de la recherche (comme l'enquête sur l'enquête de beQuali), le data paper offre une nouvelle perspective : celle d'une publication à part entière, offrant un retour sur investissement plus direct et immédiat. C'est l'occasion, pour les chercheurs, de s'interroger sur les conditions de réutilisation et la richesse

31. https://coop-ist.cirad.fr/gerer-des-donnees/rediger-un-data-paper/1-qu-est-ce-qu-un-data-paper, consulté le 28/09/2021.

de leurs données, exercice cumulatif avec des questions évoquées dans le plan de gestion de données, et ainsi de reprendre une place plus active dans ce processus. Les ingénieurs peuvent les accompagner dans cette démarche, tout en libérant un peu de temps pour d'autres tâches plus techniques associées au dépôt dans l'entrepôt (schéma de métadonnées, vocabulaires contrôlés, conversions de fichier, etc.), à moyens humains constants.

29. Une condition reste entière toutefois, si l'on veut voir se développer les *data papers*, les chercheurs et les ingénieurs doivent avoir à leur disposition un vrai modèle opérationnel – qu'ils peuvent investir plus ou moins en autonomie, et qui soit claire et optimisé du point de vue de l'ampleur du travail exigé – ainsi que des débouchés éditoriaux. Or, la situation apparaît un peu plus compliquée en pratique. Nous allons ici dire quelques mots des pistes que nous explorons actuellement pour dépasser deux écueils : la faible visibilité des *data papers* dans les communautés de recherche, en particulier dans les revues de sciences sociales, et la faible consistance de ce nouveau genre de publication.

Les *data papers* : un genre rédactionnel encore en gestation ?

30. De manière générale, les chercheurs que nous rencontrons en ont globalement peu entendu parler. Les chercheurs n'en voient quasiment pas passer dans les revues qu'ils lisent, n'en entendent pas souvent parler en séminaire, en colloque, ou dans les comités éditoriaux dont ils sont membres, et ne les voient pas mis en avant dans les critères d'évaluation de leurs établissements de tutelle. Dans les laboratoires, mis à part quelques chercheurs particulièrement sensibilisés à ce thème, ou quelques rares curieux, le sujet est principalement mis à l'ordre du jour par les ingénieurs. Plus encore, évoquer les *data papers* est parfois vécu comme revenant à ajouter de la lourdeur à l'ouverture des données de la recherche, qui est déjà souvent perçue comme une contrainte administrative. Tout l'enjeu ici consiste à faire en sorte que les *data papers* puissent être vus comme une solution à la problématique de l'ouverture des données de la recherche. Pour cela, encore faut-il avoir des références à suggérer, des modèles à montrer, sur lesquels les chercheurs vont pouvoir s'appuyer. Or, la littérature spécifiquement consacrée à discuter les contours de ce genre rédactionnel a encore assez peu circulé au-delà des spécialistes du domaine et il y a très peu de *data journals*, donc peu de débouchés pour des *data papers* (Schöpfel *et al.*, 2020)[32], et une difficulté pour les chercheurs à se représenter les attendus de ce nouveau format.

31. On trouve en effet peu de spécimens de *data papers* en sciences sociales qui pourraient servir d'exemple, en tout cas en langue française, ou du moins rédigés par

32. La consultation des bases de données bibliographiques *Web of Sciences* et *Scopus* montre que très peu de *data papers* en SHS sont publiés, et ce, même en tenant compte des biais de ces bases bibliographiques. Sur ce point, voir Le Fourner, Schöpfel, 2025.

des chercheurs et chercheuses français, auxquels les communautés disciplinaires présentes à Sciences Po pourraient davantage s'identifier. Quelques cas existent toutefois, que nous sommes régulièrement amenés à mobiliser en exemples (Dehousse *et al.*, 2017 ; Gay, 2021). La tournée de promotion de *Data Sciences Po* a également permis de repérer, au sein de Sciences Po, des pratiques de chercheurs écrivant des *data papers* « sans le savoir » (Hooghe *et al.*, 2010 ; Bakker *et al.*, 2015). Si ces papiers visent à faciliter la réutilisation des données concernées, ces premiers frémissements ne sont pas des *data papers* à part entière – les données ne sont pas déposées dans un entrepôt certifié, ni accessibles (en particulier, les questions de l'identifiant pérenne de type DOI apposé au jeu de données ou des métadonnées de description ne sont pas abordées dans les articles).

De la prospection à l'expérimentation

32. Face à cette situation, une première stratégie a consisté à trouver des débouchés éditoriaux pour les chercheurs déposants. À ce stade, nous avons trouvé un faible nombre de revues qui acceptent officiellement les soumissions de *data papers* ou de formats équivalents dans les disciplines couvertes par les dépôts. On peut supposer qu'à l'échelle de ces disciplines, comme des SHS, les tendances sont les mêmes et que ces pratiques sont très peu développées (Schöpfel *et al.*, 2020). Néanmoins, on observe que les choses évoluent : on voit par exemple que de nouvelles revues de sciences sociales prennent position en ce domaine[33], d'autres revues encourageant le dépôt de données dans des entrepôts ou demandant des *Data availability statements*[34].

33. Une deuxième stratégie a consisté à élaborer des guidelines à destination des déposants de données sur *Data Sciences Po*, afin d'aider ces derniers à préparer l'exercice de rédaction d'un *data paper*. Ces guidelines, accessibles depuis la page d'accueil de l'entrepôt, ont vocation à être actualisés en fonction des retours d'expérience que les déposants ayant été confrontés à cet exercice pourront nous faire. L'objectif est d'apporter quelques éclairages à celles et ceux qui souhaiteraient s'engager dans la rédaction de tels articles de données, autour de trois enjeux : avoir une idée claire des différences ou similarités fondamentales entre un article de recherche classique et un article de données (Le Fourner, Schöpfel, 2025), telles qu'elles apparaissent à travers les politiques éditoriales mises en place par les revues qui publient ce dernier type d'articles ; avoir une idée générale des grands types de structuration des *data papers* ; avoir des repères généraux quant aux aspects, éléments, points d'attention… à repérer, anticiper, préparer… lorsqu'on souhaite rédiger un *data paper*, qu'il s'agisse, par exemple, du temps et des compétences qu'il est nécessaire de mobiliser ou des modalités d'évaluation de ce type d'article qu'il faut

33. Par exemple, la *Revue française de sciences de l'information et de la communication* (Le Deuff 2018 ; Kembellec et Le Deuff, 2022) mais aussi d'autres revues, comme *Cybergeo*, *Humanités numériques*, ou encore, plus récemment, *L'Année sociologique*.

34. JCMS, European Law Journal, Oxford University Press (exemples tirés d'un PGD Sciences Po sur un projet de droit).

pouvoir anticiper. Une troisième stratégie a consisté à investir la création de *data journals* francophones afin de suppléer au manque d'espaces éditoriaux dédiés. Le CDSP a ainsi soutenu la création de la revue *DEMC*[35] (*Données, Méthodes, Expériences, Codes*), lancée fin 2023, qui comporte une rubrique dédiée aux *data papers*. La DRIS participe pour sa part à la revue *Data & Corpus*[36]. Le premier numéro de ce *data journal*, porté par l'université de Lorraine, est prévu pour la fin de l'année 2025.

34. Face au manque de spécimens diversifiés de *data papers*, au manque de guidelines et de retours d'expériences, nous pensons qu'il serait pertinent de mettre en commun ce travail de défrichage avec les services d'accompagnement des autres établissements de SHS. Il nous semble, à terme, nécessaire d'instaurer les conditions d'un dialogue entre producteurs de données, entrepôts de données et revues du champ francophone, pour déterminer en quoi et comment ces dernières pourraient intégrer ce nouveau type d'articles, à leur politique éditoriale (Candela *et al.*, 2015).

35. Dans cette perspective, un effet bénéfique de cette mise à plat serait de reposer une question souvent absente des réflexions sur les *data papers* : celle de leur lectorat, et derrière elle, des formes de réutilisation qui sont anticipées et donc privilégiées lorsque les jeux de données sont partagés. Il n'est pas anodin de savoir si les *data papers* sont adressés par exemple en priorité à des étudiants ayant besoin de comprendre les données qu'ils manipulent pour se former aux sciences sociales, et notamment comprendre comment les chercheurs procèdent méthodologiquement dans leurs enquêtes, ou en priorité à des chercheurs désirant faire de l'analyse secondaire. Cette question semble constituer un point aveugle. Un exemple représentatif est celui du *Research Data Journal for the Humanities and Social Sciences*[37], qui est actuellement le seul *data journal* généraliste pour les sciences sociales doté d'une assise européenne. L'examen des articles qui sont publiés dans cette revue indique que la problématique des réusages est généralement assez peu développée et, surtout, qu'elle se focalise presque exclusivement sur des visées de réanalyse à des fins de recherche, laissant de côté d'autres pans de réutilisations possibles (pour l'enseignement, ou pour l'aide à la décision publique, par exemple lorsque les données sont accessibles au-delà de la seule communauté scientifique). Cette situation est paradoxale puisque les retours d'expérience des centres de données montrent que les demandes d'usages ne se limitent pas à des finalités de recherche et intègrent, souvent massivement, des finalités alternatives, comme les finalités pédagogiques. Clarifier le lectorat cible permettrait aux chercheurs de mieux comprendre l'exercice en cernant à qui et à quels usages ils s'adressent, et donc de mieux s'approprier cette nouvelle pratique.

35. https://demc-journal.org/

36. https://www.episciences.org/fr/revues/#data-corpus

37. Notre observation porte sur 21 data papers publiés depuis 2016 (année de création de la revue) et étiquetés par la revue comme relevant des champs disciplinaires suivants : « Social and behavioural Sciences », « Social and political sciences », et « Social and Economic History ».

Le *data paper* appliqué à la biodiversité[1] : standards, outils et processus mis en œuvre pour démocratiser le concept dans la communauté de la bio-informatique

Sophie Pamerlon

Introduction

1. Le libre accès aux données et métadonnées issues de la recherche scientifique est l'un des piliers de la science ouverte et est, à ce titre, inscrit au cœur des deux plans nationaux pour la science ouverte annoncés par le ministère de l'Enseignement supérieur et de la Recherche en 2018 et 2021[2]. Outre un axe centré sur la structuration, le partage et l'ouverture des données, les deux plans s'articulent autour de deux autres axes dédiés à la généralisation de l'accès ouvert aux publications scientifiques, et à une inscription des pratiques scientifiques dans une dynamique durable, européenne et internationale.

2. Ces axes se traduisent concrètement par l'adoption de bonnes pratiques concernant notamment les réflexions sur les chaînes opératoires et structuration des données dans des plans de gestion de données (et des métadonnées qui les décrivent), leur partage – aussi bien à l'échelle nationale qu'internationale – selon les normes et standards en vigueur dans les communautés et thématiques scientifiques, leur dépôt dans des entrepôts ouverts et référencés, et leur valorisation, ainsi que la facilitation de leur réutilisation par des tiers, via des publications scientifiques – telles que les *data papers* – dans des revues à comité de lecture.

Une idée commune du GBIF et de Pensoft Publishers

3. Dans les domaines de la biodiversité et de la bio-informatique, les *data papers* ont fait leur apparition de façon sporadique au début des années 2000, dans des revues comme *Ecological Archives* ou *Earth System Science Data* (Chavan et Penev, 2011).

4. Cependant, leur publication plus généralisée, grâce notamment à la mise en place de flux, standards, outils et processus de traitement des données et des métadonnées

1. Note du directeur et de la directrice de l'ouvrage : Cette étude sort du domaine des SHS. Mais, comme il s'agit d'un modèle unique et pertinent de l'interconnexion entre entrepôts de données et plateformes de revues qui pourrait inspirer la publication de data papers en SHS, nous avons souhaité l'intégrer dans cet ouvrage.
2. Plan national pour la science ouverte https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525.

associées a véritablement commencé en 2011, suite à la décision conjointe du GBIF et des éditions Pensoft (Pensoft Publishers) de populariser la description détaillée d'un ou plusieurs jeux de données, faisant l'objet d'une publication scientifique à part entière.

5. Le GBIF (Global Biodiversity Information Facility)[3] est le Système mondial d'information sur la biodiversité. Il s'agit d'un réseau international, créé en 2001 à l'initiative du conseil scientifique de l'OCDE (Organisation de coopération et de développement économiques)[4], dont le but principal est de mettre à disposition de façon libre et gratuite les données primaires sur la biodiversité, qu'il s'agisse de données taxonomiques, de données d'occurrences d'espèces ou de données d'échantillonnage.

L'importance des métadonnées

6. Les métadonnées, ou informations descriptives sur les données, font partie intégrante du modèle de données du GBIF depuis sa création[5], et ne sauraient être dissociées des jeux de données qu'elles décrivent et avec lesquels elles sont systématiquement publiées sur le site GBIF.org ; elles apparaissent sur la page de chaque jeu de données mis en ligne sur le site afin de faciliter son interprétation et sa potentielle réutilisation par de tierces personnes (figure 1).

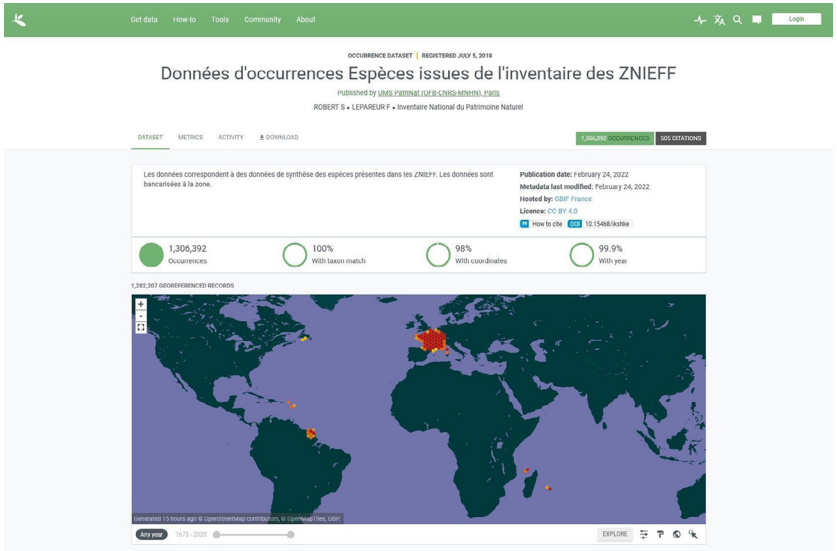


Figure 1. Page d'un jeu de données sur le site GBIF.org avec ses métadonnées descriptives[6]

Crédit : OpenStreetMap contributors, OpenMapTiles, GBIF

7. La rédaction, la mise à jour et l'enrichissement des métadonnées constituent un processus parfois complexe et chronophage, mais nécessaire dans le cycle de vie de la donnée et des jeux de données. La production de métadonnées complètes et détaillées offre de nombreux avantages : une meilleure visibilité

3. GBIF https://www.gbif.org/fr/.
4. OCDE https://www.oecd.org/fr/.
5. GBIF data standards https://www.gbif.org/standards.
6. Jeu de données : Robert S., Lepareur F., Inventaire National du Patrimoine Naturel (2022). Données d'occurrences Espèces issues de l'inventaire des ZNIEFF. Version 1.7. UMS PatriNat (OFB-CNRS-MNHN), Paris. Occurrence dataset https://doi.org/10.15468/ikshke accessed via GBIF.org on 2022-03-15.

des données à l'échelle nationale et internationale, la mise en valeur du temps et des ressources consacrées à leur collecte et à leur gestion, et la reconnaissance des différents acteurs impliqués à toutes les étapes de leur traitement (Chavan et Penev, 2011).

Le rôle d'un *data paper*

8. Afin de valoriser ce travail, le GBIF s'est associé à l'éditeur bulgare Pensoft Publishers[7] pour proposer un premier modèle de rédaction semi-automatisée de *data paper* dès 2011, modèle supporté par les outils GBIF déjà en place et les standards utilisés par la communauté internationale autour des données de biodiversité. Cette première étape a ensuite été complétée par la mise en œuvre d'interfaces de rédaction et d'édition par Pensoft Publishers, et d'autres maisons d'édition scientifiques, dans les années suivantes.

9. Le *data paper*, comme décrit par Chavan et Penev en 2011 (après une première ébauche en 2009, cf. Chavan et Ingwersen, 2009), a ainsi pour objectif principal de décrire finement un ou plusieurs jeux de données, et non des résultats de travaux de recherche, tout en informant la communauté scientifique de la disponibilité de ces jeux de données et de leur potentiel pour des utilisations futures (Dedieu, 2014).

10. Il permet également de mettre en valeur les équipes de recherche ou d'expertise (et notamment les personnels techniques peu souvent cités dans les publications scientifiques habituelles), ainsi que le temps passé par ces personnes à collecter, informatiser, numériser, traiter, conserver, formater, maintenir et mettre en ligne les données de biodiversité.

11. Tout jeu de données ou ensemble de jeu de données peut potentiellement faire l'objet d'un *data paper*, quels que soient son *scope*, protocole ou type de données ; dans le domaine de la biodiversité, les exemples de *data papers* publiés concernent aussi bien les sciences participatives que les systèmes d'agrégation de données naturalistes, les jeux de données de la recherche scientifique, le suivi des populations ou encore les données issues de zones naturelles d'intérêt.

12. Leur point commun est de décrire de façon précise et factuelle les systèmes, protocoles, méthodes de collecte ou d'observation des données, mais également leur informatisation, numérisation, agrégation ou les protocoles de traitement qui leur sont appliqués avant leur partage sur des entrepôts ou portails librement accessibles tels que le site GBIF.org ; plus généralement, le but d'un *data paper* est de faire état de l'originalité et du potentiel de réutilisation des données décrites par d'autres scientifiques (Dedieu, 2014).

13. Les jeux de données ainsi décrits dans une revue à comité de lecture sont donc plus visibles, accessibles et

7. Pensoft https://pensoft.net/.

facilement réutilisables par de tierces personnes, sans qu'elles aient besoin de requérir des informations supplémentaires auprès des producteurs des données (figure 2).

Figure 2. Exemple d'un *data paper* (Lepareur *et al.*, 2022)

La chaîne éditoriale

14. L'un des prérequis obligatoires pour la soumission et l'éventuelle publication d'un *data paper* est la mise en ligne préalable des données décrites, de façon libre et gratuite, sur un entrepôt de données ou un portail de consultation et téléchargement des données de biodiversité tel que le site GBIF.org.

15. Il est également possible d'envoyer les données faisant l'objet d'un *data paper* directement à la maison d'édition comme matériel annexe, mais cette option n'est pas recommandée, car les données mises à disposition par ce biais sont moins facilement trouvables et interopérables, contrairement aux deux premières options ; comme d'autres entrepôts liés à la science ouverte, le GBIF garantit notamment la mise en œuvre des principes FAIR (Wilkinson *et al.*, 2016) en faisant en sorte que les données et les métadonnées associées soient Faciles à trouver, Accessibles, Interopérables et Réutilisables (figure 3), ce qui représente un avantage certain pour les institutions ou autres éditeurs de données qui les partagent de cette façon.

Attribution	Digital Object Identifiers assignés aux : • Jeux de données publiés • Téléchargements des utilisateurs • Data papers
Licences	Les éditeurs choisissent une licence appropriée: • CC-0 • CC-BY • CC-BY-NC
Standards	Données: DarwinCore, ABCD, DublinCore Metadonnées: Ecological Metadata Language (EML)
Accessibilité	GBIF.org portail d'accès aux données • Requêtes interopérables • Cartes interactives • Téléchargements gratuits • web services
Publication	Integrated Publishing Toolkit (IPT), BioCASE, Spreadsheet templates

Figure 3. Récapitulatif des différents outils et processus FAIR mis en place par le GBIF

Crédit : GBIF

16. Les données décrites dans un *data paper* peuvent ainsi être publiées en amont dans le GBIF via l'outil Integrated Publishing Toolkit (IPT)[8] qui les formate et les envoie vers le portail GBIF.org sous la forme d'archives Darwin Core (DwC-A)[9] contenant les données et les métadonnées associées.

17. Le GBIF indexe et affiche ensuite en ligne ces données et leur description, qu'il est possible de récupérer via l'IPT dans un format compatible avec l'interface de soumission Pensoft pour les *data papers*.

18. Il est ensuite possible de modifier, compléter et illustrer le *data paper* sur l'interface de l'outil d'aide à la rédaction proposé par Pensoft Publishers, ou son équivalent chez de nombreux autres journaux scientifiques, avant de le soumettre à la revue choisie, qui applique par la suite le processus habituel de relecture par les pairs et d'édition classique comme tout autre article scientifique.

19. Ce « peer-reviewing » inclut notamment l'analyse des données décrites dans le *data paper*, en plus du texte de l'article en lui-même, afin de s'assurer de la qualité, de la cohérence, du formatage et de l'accessibilité des données dont il fait l'objet.

20. Pour mieux faciliter la citation et le suivi des articles et des données, un système d'identification unique des *data papers* et des jeux de données associés a également été adopté par le GBIF et de nombreuses maisons d'édition, le Digital Object Identifier (DOI).

21. Dans le cadre de la collaboration entre GBIF et Pensoft, les bonnes pratiques recommandent de citer dans chaque *data paper* le DOI du jeu de données décrit dans l'article, et inversement, d'éditer les métadonnées associées à ce jeu de données sur le site du GBIF une fois le *data paper* publié, pour y faire figurer le DOI de l'article (Chavan et Penev, 2011).

22. Le versionnage des jeux de données est également pris en compte : dans le cas d'un jeu de données évolutif dont le nombre ou l'intégrité des données sont voués à changer au fil du temps et des mises à jour, il est recommandé de citer non seulement son DOI, mais aussi son numéro de version afin de garantir l'accès à la version du jeu de données décrite dans le *data paper*, et non une version ultérieure potentiellement modifiée.

Un premier bilan

23. Depuis le lancement du processus, plus de 430 *data papers* ont été publiés par la communauté internationale travaillant sur les données primaires de biodiversité[10], sans compter ceux parus dans d'autres domaines de la recherche scientifique (médecine, physique, sciences

8. IPT https://www.gbif.org/fr/ipt.
9. Darwin Core https://ipt.gbif.org/manual/en/ipt/2.5/dwca-guide.
10. Cf. https://www.gbif.org/resource/search?contentType=literature&topics=DATA_PAPER&relevance=GBIF_PUBLISHED.

environnementales, mais aussi sciences humaines et sociales).

24. Il est à noter que la publication d'un *data paper* n'est pas réservée aux seuls acteurs de la recherche ; plusieurs exemples récents montrent un intérêt croissant des communautés d'acteurs de l'expertise et des politiques publiques pour les *data papers*, qui peuvent aussi décrire des données issues de programmes publics ou privés hors de la sphère scientifique (cf. Ichter *et al.*, 2022 ; Lepareur *et al.*, 2022).

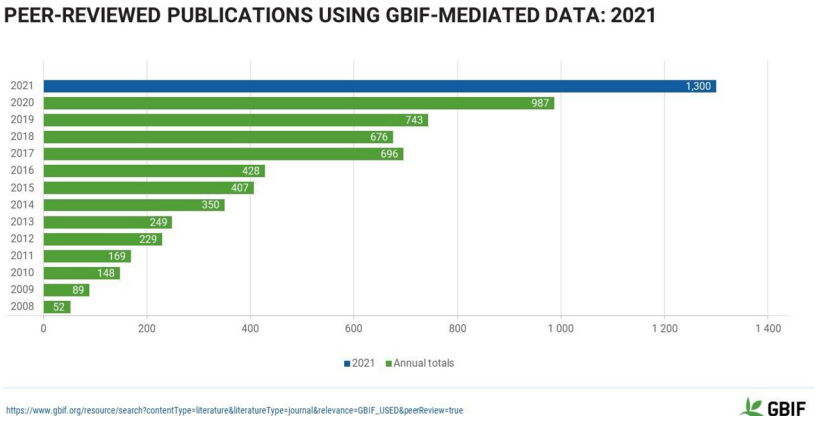


Figure 4. Évolution par année du nombre de publications scientifiques citant des données partagées sur GBIF.org
Crédit : GBIF

25. Sur le site du GBIF, il est possible de retrouver la liste des *data papers* publiés par pays, année ou sur un territoire donné ; en plus des *data papers*, la liste complète des publications scientifiques basée pour tout ou partie sur des données diffusées via le GBIF est accessible dans la revue scientifique[11] annuelle du réseau (*GBIF Science Review*, figure 4).

26. L'établissement et la durabilité d'un concept tel que le *data paper* requièrent de solides bases techniques au sein des communautés concernées, et notamment l'adoption d'outils et de standards communs pour optimiser le partage et l'interopérabilité des données et des métadonnées. Des mécanismes concrets peuvent être mis en place pour faciliter le plus possible les étapes de rédaction et de soumission des *data papers*, comme cela a été fait par le GBIF et Pensoft Publishers ces dix dernières années, afin de fournir un soutien technique fiable et évolutif aux producteurs de données et aux auteurs de *data papers*.

27. Le *data paper* n'est pas un concurrent des articles scientifiques habituellement publiés dans le cadre de projets de recherche, mais bien un complément d'informations qui peut être publié en parallèle d'un ou plusieurs autres articles consacrés à l'analyse des résultats de recherche, ou des données collectées plus généralement.

28. Là encore, un suivi et une liaison entre les divers articles et jeux de données peuvent être effectués via l'utilisation d'identifiants uniques tels que les DOI attribués à chaque ressource en ligne (figure 5).

11. GBIF Secretariat, « Science Review ».

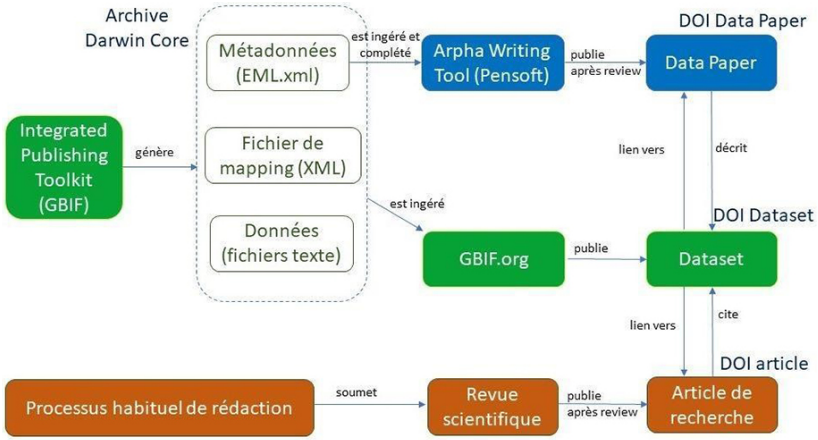


Figure 5. Schéma récapitulatif du processus de publication en parallèle d'un jeu de données, du *data paper* le décrivant, et d'un article scientifique analysant ces données

Crédit : Sophie Pamerlon, PatriNat (CNRS-OFB-MNHN)

29. Par la suite, nous allons décrire plus en détail deux standards utilisés par ce workflow ainsi que les deux outils développés pour la génération et la publication des métadonnées (IPT) et pour la rédaction d'un *data paper* (Arpha Writing Tool).

Standards

Ecological Metadata Language (EML)

30. Au sein du GBIF, et plus généralement de la communauté travaillant sur les données de biodiversité, la mise en forme et l'échange de métadonnées décrivant les jeux de données primaires sur la biodiversité s'effectuent principalement grâce au standard Ecological Metadata Language (EML)[12].

31. Ce standard, développé en 1997 par l'université de Californie et maintenu par des contributeurs bénévoles jusqu'à aujourd'hui, est une spécification permettant une description fine des métadonnées, qui peuvent ainsi être facilement interprétées par les personnes souhaitant consulter ou réutiliser les données de biodiversité auxquelles les métadonnées sont associées.

32. Les bonnes pratiques scientifiques, *a fortiori* dans le cadre de la rédaction d'un *data paper*, recommandent de détailler le plus précisément possible les métadonnées, afin d'éviter toute ambiguïté sur le contenu du jeu de données qu'elles décrivent et ainsi en permettre une réutilisation et une interprétation correctes.

33. Comme indiqué sur le site officiel du standard, l'EML est défini « par un ensemble de schémas XML définissant les types et la structure d'un document EML valide » ; il s'agit ainsi d'un standard facilement interprétable par des machines ou logiciels, tels que ceux utilisés par le GBIF et Pensoft Publishers pour produire, importer et soumettre des manuscrits de *data papers* à partir des métadonnées de jeux de données.

12. EML https://eml.ecoinformatics.org/.

GBIF Metadata Profile (GMP)

34. Dans le réseau GBIF, décision a été prise lors des premiers développements informatiques pour la communauté d'utiliser une sous-catégorie de l'EML, choisie spécifiquement afin de standardiser la description des jeux de données sur le portail GBIF : le GBIF Metadata Profile (GMP)[13].

35. Ce standard est principalement basé sur la version 2.1 de l'EML (le passage à la version 2.2 étant prévu au printemps 2023), et ne conserve que les balises EML adéquates pour correspondre aux besoins de la communauté GBIF et aux outils développés pour partager les données, tout en y ajoutant des champs supplémentaires tels que la licence attribuée au jeu de données (à choisir parmi trois options de licences ouvertes Creative Commons : CC0, CC-BY et CC-BY-NC)[14], ou des champs de métadonnées additionnelles non présents dans l'EML (URL du logo institutionnel, fréquence de mise à jour, informations diverses sur le processus de traitement des données, etc.).

36. Ce profil de métadonnées GMP est compatible avec d'autres formats de métadonnées communément utilisés, tels que le profil de métadonnées ISO 19139[15], et est associé à chaque jeu de données publié dans le GBIF sous la forme d'un fichier XML facilement interprétable.

13. GMP https://github.com/gbif/gbif-metadata-profile.
14. Cf. https://www.gbif.org/news/82363/new-approaches-to-data-licensing-and-endorsement.
15. ISO/TS 19139-1:2019 https://www.iso.org/fr/standard/67253.html.

Outils

Integrated Publishing Toolkit (IPT)

37. Le GBIF Metadata Profile est majoritairement utilisé au travers de l'Integrated Publishing Toolkit (IPT), un logiciel libre et gratuit écrit en Java et développé par le GBIF, afin de publier et de partager des jeux de données de biodiversité dans le réseau GBIF. L'IPT peut aussi être associé à un compte DataCite ou EZID[16] de façon à associer un DOI à un jeu de données, afin de créer un entrepôt de données local.

38. Le but principal de l'IPT est de permettre aux institutions productrices et détentrices de données sur la biodiversité de les publier aisément, en utilisant les standards d'échange de données et métadonnées communément adoptés dans le réseau GBIF. Pour cela, le logiciel IPT propose une interface utilisateur simple via laquelle il est possible de charger un ou plusieurs fichiers sources de données (ou un fichier déjà en ligne de type document partagé), ou de connecter le logiciel directement sur une base de données.

39. Une fois publiée via l'IPT, la ressource est visible sur la page d'accueil du logiciel (figure 6), et sur le site du GBIF si elle a été préalablement liée à une institution membre du GBIF lors de sa mise en ligne.

16. EZID https://ezid.cdlib.org/.

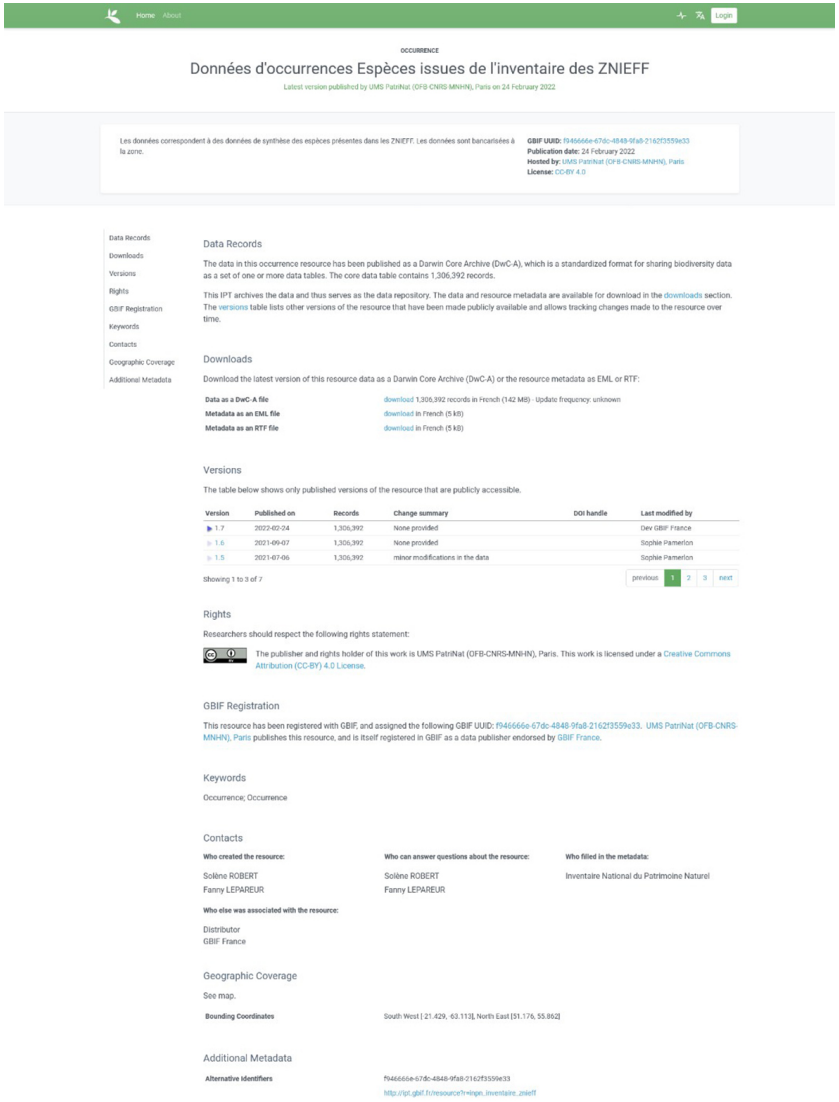


Figure 6. Page publique d'une ressource (jeu de données de biodiversité) publiée sur une installation IPT, avec ses métadonnées descriptives

40. La page publique de la ressource sur l'IPT donne accès aux données formatées au standard Darwin Core, aux métadonnées et à d'autres informations complémentaires (lien vers la page du jeu de données sur GBIF.org, versions, institution détentrice des données).

41. Il suffit ensuite d'effectuer le « mapping » (ou mise en correspondance des champs) avec le standard Darwin Core utilisé par la communauté GBIF et ses partenaires pour partager les données de biodiversité. Ce mapping est l'étape consistant à faire correspondre chaque nom de champ (colonne) du fichier source avec un des termes du standard Darwin Core, pour une meilleure interopérabilité des données par la suite.

42. Cela permet de partager plusieurs types de jeux de données vers le GBIF et ses partenaires : jeux de données d'occurrences d'espèces, listes taxonomiques, ou événements d'échantillonnage (jeux de données plus complexes issus de protocoles d'observation ou de collecte spécifiques).

43. Il est également possible de publier vers le GBIF des jeux de données « metadata-only », c'est-à-dire uniquement constitués d'un fichier de métadonnées sans données associées. Cette option est particulièrement utile afin de faire connaître l'existence de certains jeux de données non publiés pour diverses raisons : collections scientifiques non informatisées ou données liées à des articles scientifiques encore sous embargo, par exemple. Elle n'est cependant pas adaptée à la publication d'un *data paper*,

qui exige la mise en ligne préalable des données décrites sur un entrepôt d'accès libre et ouvert.

44. Afin d'achever la mise en ligne de données via l'IPT, la dernière étape consiste à compléter les métadonnées au format GMP (basé sur l'EML), avant de publier le tout sous forme d'archive Darwin Core (DwC-A). Depuis 2021, il est également possible de charger un fichier EML déjà prêt directement sur l'IPT, afin d'éviter une saisie manuelle des métadonnées.

45. L'archive Darwin Core (figure 7) est le format de partage des données et des métadonnées associées, généré automatiquement par l'IPT lors de la publication d'une ressource, et prend la forme d'un dossier zippé contenant les données au format texte (.txt), le document de correspondance des champs (fichier XML) et le document de métadonnées, également au format XML.

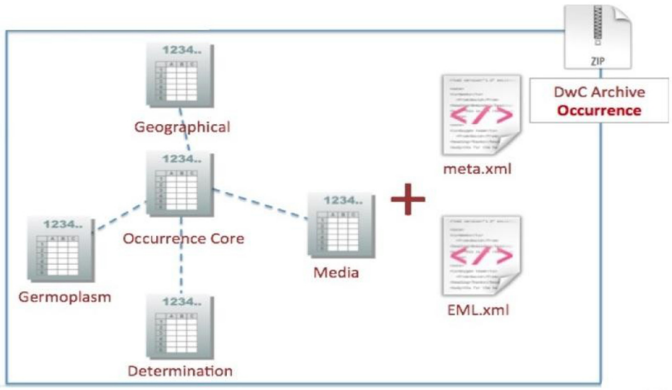


Figure 7. Schéma d'une archive Darwin Core
Crédit : GBIF

46. Une archive Darwin Core générée par une installation IPT contient les fichiers de données (ici un cœur de standard « Occurrence Core » connecté en étoile avec les fichiers d'extensions), le document de correspondance des champs (meta.xml) et le document de métadonnées (eml.xml).

47. Cette archive peut ensuite être moissonnée par le portail GBIF après activation du bouton « Register » de l'IPT pour que son contenu soit visible, accessible, consultable et téléchargeable en ligne sur GBIF.org.

48. Dès le lancement du processus de publication des données, une page dédiée à chaque jeu de données est créée automatiquement sur le portail GBIF.org, après indexation et analyse des données par le GBIF pour les interpréter et signaler de potentiels soucis de taxonomie ou de géoréférencement via des drapeaux (« flags ») visibles sur la page de chaque occurrence.

49. Un DOI est attribué par le GBIF à chaque jeu de données dès sa publication sur GBIF.org, sauf dans le cas où le jeu de données en possède déjà un ; le DOI d'origine peut alors être repris par le GBIF pour éviter les confusions et faciliter le traçage de cette ressource.

50. La publication d'un jeu de données dans le réseau GBIF (ou simplement sur l'IPT même avant moissonnage par le GBIF) n'est pas possible s'il n'y a pas de fichier de métadonnées associé à ce jeu de données.

51. Sur l'IPT, les métadonnées d'une ressource (jeu de données) déjà publiées sont disponibles à la consultation et au téléchargement sur la page de cette ressource (figure 8), sous deux formats : RTF (Rich Text Format, pour une lecture ou édition accessible par un être humain via n'importe quel outil de traitement de texte), ou EML (fichier XML). Les deux fichiers sont automatiquement générés par l'IPT à partir des métadonnées remplies via l'interface utilisateur, et peuvent être modifiés à volonté, puis être pris en charge par d'autres outils, tels que l'Arpha Writing Tool de Pensoft Publishers (cf. plus loin).

Figure 8. Détail de la page publique d'une ressource publiée sur une installation IPT

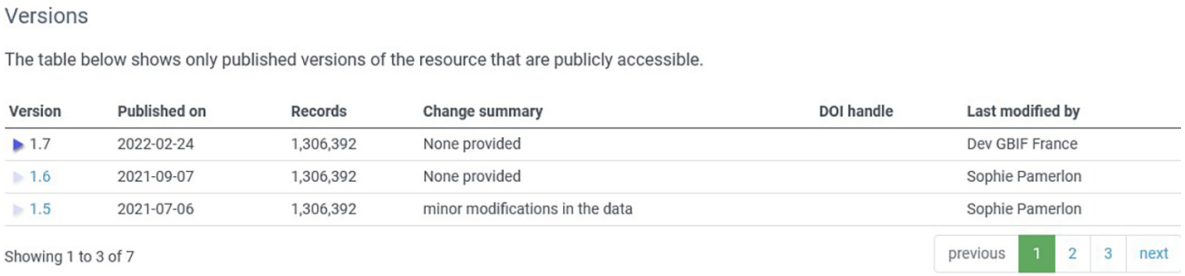


Figure 9. Détail de la page publique d'une ressource publiée sur une installation IPT

52. Sur la page publique d'une ressource se trouvent les différentes options de téléchargement des données (archive Darwin Core) et métadonnées (format EML et RTF pour les métadonnées).

53. La ou les personnes en charge de la publication d'une ressource via l'IPT peuvent à tout moment accéder à la ressource en se connectant sur l'IPT et modifier les données, les métadonnées ou le « mapping », en republiant la ressource ensuite afin d'enregistrer les modifications effectuées.

54. Cela génère automatiquement une nouvelle version de la ressource (figure 9) ; les versions précédentes peuvent être sauvegardées et accessibles sur l'IPT si le mode archivage est activé. Une ressource éditée sur l'IPT est également mise à jour côté GBIF si elle était déjà publiée sur le portail GBIF.org.

55. Il est possible d'accéder aux versions antérieures d'un jeu de données si l'option « archival mode » de l'IPT est activée.

56. Pour publier un jeu de données vers le GBIF sous la forme d'une archive Darwin Core, seule la première page des métadonnées (« Basic metadata ») doit absolument être complétée via l'IPT ou un autre système de création et d'envoi d'archives Darwin Core (figure 10) ; il est

cependant recommandé de remplir le plus de champs possible dans les métadonnées, afin de décrire le jeu de données de façon précise et exhaustive et ainsi éviter de potentielles erreurs d'interprétation des données ou de leur contexte de collecte ou d'observation.

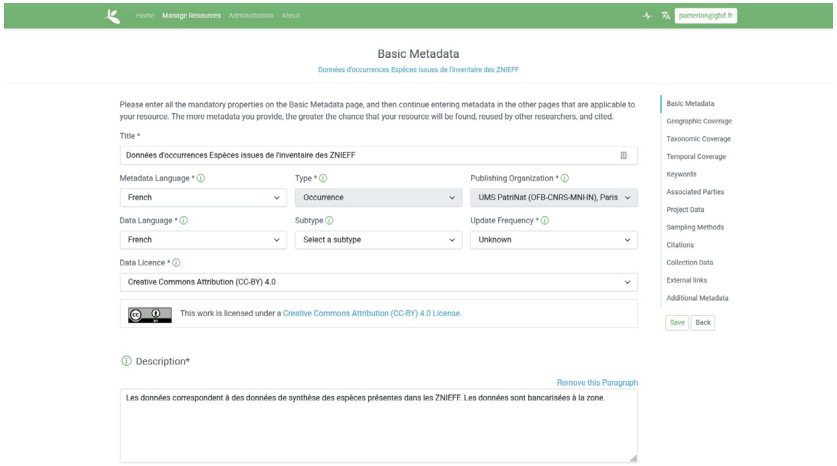


Figure 10. Détail de la section « Métadonnées » d'une ressource publiée sur une installation IPT

57. La section « Métadonnées » permet au(x) gestionnaire(s) d'un jeu de données d'éditer les différentes catégories de métadonnées apparaissant sur la droite de la page, après connexion à l'IPT.

58. Les métadonnées de base (« Basic metadata ») contiennent les éléments de description essentiels d'un jeu de données : titre, nom de l'institution détentrice des données, licence attribuée au jeu de données, description succincte, fréquence de mise à jour des données si nécessaire, informations de contact.

59. Les autres parties de la section « Métadonnées » sont facultatives, mais très utiles à compléter et permettent de préciser plus d'éléments contextuels relatifs au jeu de données dans son ensemble : couverture géographique, couverture taxonomique, couverture temporelle, mots-clefs, contacts additionnels (contributeurs d'un projet de science participative, par exemple), informations sur le projet dans le cas de données collectées dans le cadre d'un projet ou programme particulier (identifiant du projet, financeurs, personnel lié au projet (Principal Investigator), protocole détaillé de collecte ou d'observation des données (pièges photo, stations de pêche, transects, points d'écoute ornithologiques, etc.), citations et bibliographie, liens divers et autres métadonnées additionnelles (logo, autres identifiants du jeu de données, processus de nettoyage des données, etc.).

60. On retrouve tous ces éléments (obligatoires comme facultatifs si ces derniers sont complétés) dans les fichiers EML et RTF présents sur l'IPT, ainsi que sur la page du jeu de données sur le site du GBIF et lors du téléchargement des données depuis ce même site.

61. Ces informations contribuent à former le « squelette » du *data paper*, chaque catégorie de métadonnées étant ensuite interprétée et reconnue par les outils d'aide à la rédaction des maisons d'édition scientifique comme une composante du *data paper* ; d'autres catégories hors EML (Résumé, Informations nouvelles, Traits d'espèces…) peuvent être proposées par les outils d'aide à la rédaction, comme l'Arpha Writing Tool décrit ci-après.

Arpha Writing Tool de Pensoft

62. Pensoft Publishers est une maison d'édition de littérature scientifique, fondée en 1992 en Bulgarie et proposant plusieurs revues (majoritairement liées à des disciplines de la biodiversité et de l'environnement) en accès ouvert[17].

63. Partenaire institutionnel du réseau GBIF depuis 2010, Pensoft Publishers contribue, via l'enregistrement de quatre de ses revues à comité de lecture en tant que « data publisher » (éditeur de données) dans le réseau GBIF, à partager à l'échelle internationale plusieurs dizaines de milliers d'occurrences d'espèces publiées dans des articles scientifiques.

64. C'est dans ce contexte que le GBIF et Pensoft se sont associés en 2011 afin de mettre en place un processus rapide et efficace pour générer un fichier EML complet et interprétable à toutes les étapes du processus de publication du *data paper*.

65. Historiquement, il existe deux options possibles pour soumettre un *data paper* : l'envoi manuel du manuscrit au comité de lecture de la revue choisie, ou le chargement d'un fichier EML déjà généré en amont, qui est ensuite interprété par un des outils d'aide à la publication développés par de nombreuses maisons d'édition scientifique à l'heure actuelle.

66. Comme mentionné précédemment, l'IPT du GBIF propose l'export des métadonnées sous deux formats : RTF, compatible avec la plupart des logiciels de traitement de texte actuels, ou EML (fichier XML « machine-readable », c'est-à-dire interprétable par une machine ou logiciel dédié).

67. Cela correspond aux deux voies possibles proposées par Pensoft et d'autres éditeurs scientifiques, qui acceptent aussi bien les manuscrits rédigés, complétés ou corrigés via un éditeur de texte, que des fichiers EML modifiables par la suite sur l'interface de l'outil d'aide à la rédaction (figure 11).

Figure 11. Résumé du processus de publication d'un *data paper* dans une revue Pensoft Publishers à partir d'un document de métadonnées EML

Crédit : Pensoft Publishers

17. Pensoft https://pensoft.net/.

68. Fichier RTF ou XML reprennent toutes les sections de métadonnées complétées via l'interface de l'IPT ; après ingestion et analyse par l'outil d'aide à la rédaction, tel que Arpha Writing Tool, chacune de ces sections descriptives peut ensuite être enrichie par des compléments d'information, des figures, cartes et autres illustrations, ou des données et éléments annexes.

69. L'outil Arpha (Arpha Writing Tool), développé par Pensoft Publishers, est une interface de rédaction, édition et soumission d'articles scientifiques associées aux revues à comité de lecture de la maison d'édition.

70. Mis en place dans l'optique d'offrir un support de rédaction simple aux auteurs, il propose plusieurs fonctionnalités pour créer et gérer un manuscrit d'article (*data paper* inclus) : il est en effet possible d'envoyer un manuscrit (document de métadonnées au format RTF par exemple) à l'éditeur pour que celui-ci le mette en forme avant soumission, ou de créer directement un nouvel article sur la plateforme via l'importation d'un fichier de type XML ou autre format couramment utilisé dans les archives et entrepôts de données ; l'import par webservice ou API (Application Programming Interface) est également possible et permet de charger n'importe quel format de document de métadonnées.

71. L'outil propose également la possibilité de rédiger le *data paper* directement via son interface, sans charger de fichiers de métadonnées.

72. Outre le workflow semi-automatisé de production de métadonnées mis en place en collaboration avec le GBIF, l'outil d'aide à la rédaction Arpha est également compatible avec l'entrepôt de données Dryad Data Repository et d'autres entrepôts thématiques, tels que (entre autres) le Consortium for Barcode of Life (CBOL), GenBank, ou Pangea.

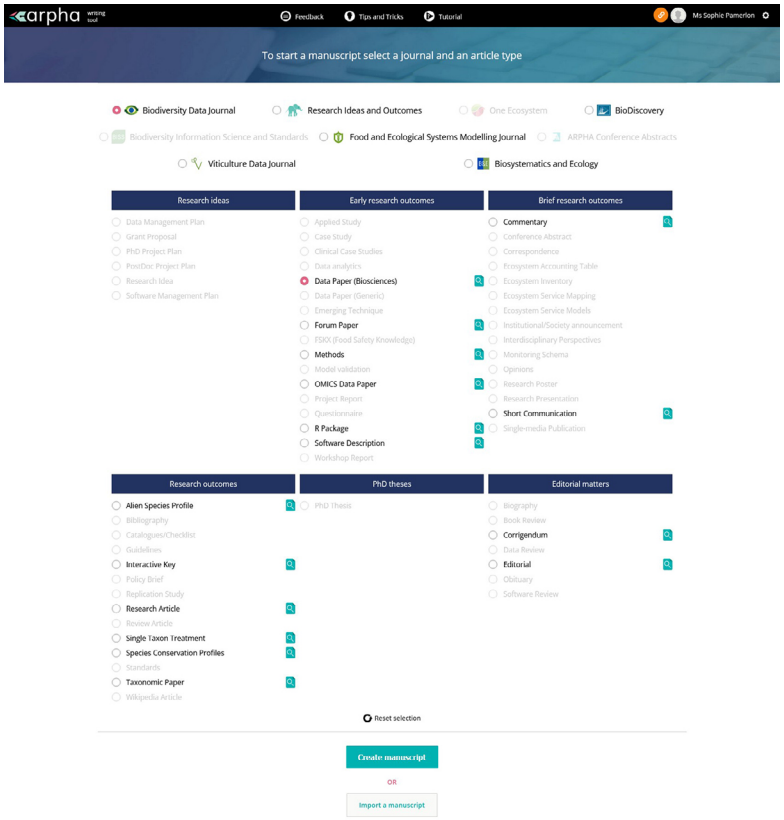


Figure 12. Capture d'écran de la page Arpha Writing Tool permettant de choisir une revue dans laquelle soumettre un *data paper*

73. Dans tous les cas, la création d'un manuscrit de *data paper* via cette interface nécessite de se connecter à l'outil ou de créer un compte, idéalement un par auteur (chaque auteur pourra ensuite être invité à éditer le manuscrit sur l'interface utilisateur), puis de choisir le journal le plus adéquat auquel le soumettre (figure 12).

74. Dans le cas de Pensoft Publishers, c'est principalement la revue *Biodiversity Data Journal*[18] qui est privilégiée pour l'édition de *data papers*, l'interface de l'outil permettant ensuite le chargement d'un fichier de métadonnées (figure 13).

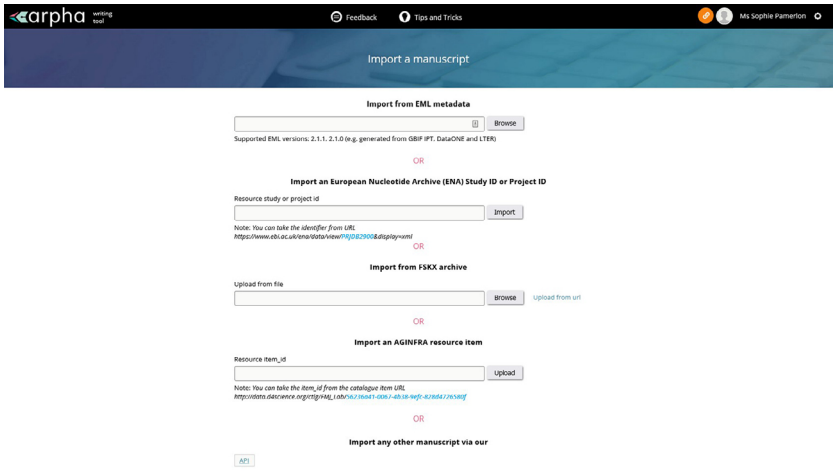


Figure 13. Capture d'écran de la page Arpha Writing Tool permettant de charger un document de métadonnées EML (ou autre format accepté par l'outil) afin de créer un *data paper*

18. *Biodiversity Data Journal* https://bdj.pensoft.net/.

75. Une fois le manuscrit envoyé, importé ou créé directement sur l'interface de Arpha Writing Tool, il est possible de compléter ou d'éditer chaque composante des métadonnées : titre, informations de contact, résumé, description, et autres catégories des métadonnées (figure 14).

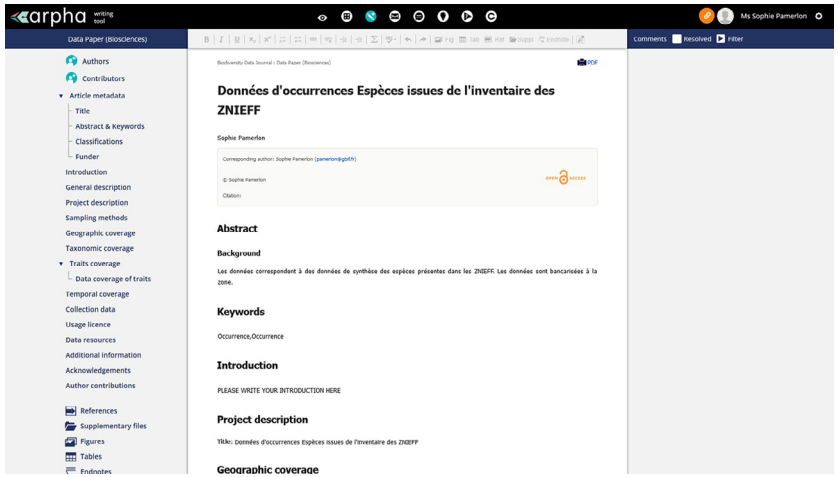


Figure 14. Capture d'écran de la page de rédaction d'un article ou *data paper* sur Arpha Writing Too, après chargement d'un document de métadonnées existant

76. Le suivi des modifications est effectué via un bouton dédié en haut de page ; un tableau de bord accessible via le profil de chaque utilisateur propose également un suivi de chaque article publié ou manuscrit en cours de rédaction.

77. L'interface permet d'ajouter, via l'envoi d'un lien d'invitation, autant d'auteurs ou contributeurs que nécessaire,

lesquels peuvent ensuite éditer leur profil pour y ajouter leur identifiant de type ORCID ou assimilé. Il est également possible de faire appel à des relecteurs externes avant la soumission finale, pour détecter oublis ou incohérences qui auraient échappé à la vigilance des auteurs.

78. Une fois prêt, le manuscrit doit passer une validation automatique par l'outil Arpha (bouton « *Validate* » en bas de page), qui détecte incohérences techniques et omissions dans le *data paper* (champs manquants, figures non citées dans le corps du texte, etc.). Le bouton « *Submit for technical review* » permet ensuite d'envoyer le manuscrit et ses figures ou données annexes à un évaluateur technique, qui émet conseils et préconisations pour assurer le bon formatage des données et la cohérence générale du texte. Cette revue technique nécessite souvent des corrections dans les fichiers descriptifs des données ou dans les figures associées au *data paper*.

79. Après validation et revue technique, le manuscrit de *data paper* peut être officiellement soumis à la revue choisie grâce au bouton « *Submit* », qui n'apparaît qu'à cette étape. Il est alors soumis à l'évaluation par un comité de lecture, au même titre qu'un article scientifique habituel ; il est utile de garder en mémoire que, bien que la publication d'un *data paper* soit soumise à moins d'enjeux, son acceptation par la revue choisie n'est en aucun cas garantie.

80. Une fois le *data paper* accepté après revue technique et comité de lecture, il apparaît sur le site de la revue ou du journal choisi comme tout autre article scientifique. Grâce au DOI qui lui est attribué dès sa parution, ses citations potentielles dans d'autres articles sont suivies via un onglet ou section dédiée (par exemple l'onglet « *Cited* » sur la page d'un *data paper* sorti dans une revue Pensoft, ou la section « *This article is cited by* » dans la revue *Nature*).

81. Pour plus d'informations sur les *data papers* liés aux données de biodiversité, le site du GBIF propose une page dédiée compilant les *data papers* déjà publiés au sujet de jeux de données déjà déposés dans le GBIF, une liste évolutive de revues scientifiques et techniques acceptant la publication de *data papers* (avec leur facteur d'impact et frais de publication respectifs), et d'autres informations sur le sujet[19].

82. Il est également possible de consulter, via le portail GBIF.org ou les sites des journaux Pensoft, les *data papers* déjà publiés ou en pré-print, grâce à des filtres permettant de trier les articles par pays, auteur, date ou mots-clefs[20].

83. Le site du point nodal GBIF France propose également une page dédiée aux *data papers* publiés dont au moins l'un des auteurs est rattaché à un organisme de recherche ou d'expertise français, et l'équipe GBIF France fournit un appui technique concernant tous les aspects de la mise

19. GBIF data papers https://www.gbif.org/data-papers.
20. Cf. note 8.

en ligne des données et métadonnées, ou de la rédaction et soumission d'un *data paper*.

Conclusion

84. De par sa nature FAIR et sa validation par un comité de lecture, le *data paper* est devenu, dès son origine, un élément essentiel de la science ouverte et plus particulièrement des bonnes pratiques liées à l'accessibilité, valorisation et réutilisation des données.

85. Bien qu'il ne soit pas obligatoire, il est de plus en plus fréquemment intégré dans les plans de gestion de données, et devient fortement recommandé comme livrable en parallèle du dépôt de données lors de financement de projets. Le *data paper* permet en effet d'assurer une description structurée, exhaustive et à jour des données de recherche par leurs propres auteurs pour leur compréhension et réutilisation par des tiers, facilitant notamment la reproductibilité de la Recherche et la mise en valeur de données, parfois peu connues ou accessibles avant leur ouverture et leur description dans un document aisément interprétable.

86. Cette valorisation des données et métadonnées dans un format structuré facilite de plus les échanges entre chercheurs, personnels techniques ou d'accompagnement et autres gestionnaires de données, et ouvre ainsi la voie à des échanges fructueux et de nouvelles pistes de recherches dans tous les domaines scientifiques et techniques, ainsi qu'à leurs potentielles mises en application dans le cadre de politiques publiques.

87. S'il existe encore peu d'analyses chiffrées de l'impact des *data papers* sur la production scientifique dans les domaines des « sciences dures », une étude à ce sujet dans le domaine des sciences humaines et sociales a montré un impact significatif et positif de l'ouverture des données et des *data papers* dans les disciplines concernées, notamment pour maximiser la réutilisation des données (McGillivray *et al.*, 2022).

88. Dans le domaine de la biodiversité et plus spécifiquement en France, la plupart des retours d'expérience concernant les *data papers* font état d'articles plus simples et rapides à publier que les articles scientifiques classiques consacrés à des analyses ou résultats de recherches (comm. pers.). Le soutien des équipes et reviewers techniques, associés aux flux de données et de métadonnées déjà mis en place entre entrepôts et outils d'aide à la rédaction, ainsi que l'utilisation de standards internationaux pour en faciliter la structuration, expliquent ces délais raccourcis et la relative facilité de publication de ce genre d'article.

89. Ces avantages contribuent à faire du *data paper* un outil indispensable de la science ouverte, dans le domaine de la biodiversité et de la bio-informatique, comme dans d'autres thématiques scientifiques.

Perspectives

Christine Kosmopoulos, Victoria Le Fourner et Joachim Schöpfel

Un écosystème en émergence

1. D'un point de vue purement quantitatif, les *data papers* ne représentent pas en l'état actuel un volume significatif dans le paysage de la publication scientifique en général, ni même en SHS. S'agit-il pour autant d'un simple épiphénomène ? Nous ne le pensons pas. En effet, non seulement les *data papers* font partie intégrante de la bibliodiversité de la recherche, mais, en plus, et surtout, ils sont un élément central de l'écosystème reliant données et publications, une sorte de *missing link*, l'articulation manquante entre les entrepôts de données et les plateformes de revues.

2. L'écosystème des données est un système dynamique en émergence, dans lequel la diversité est de règle. Ni ses éléments constitutifs – plateformes, entrepôts, revues, articles, données, formats, standards, identifiants, etc. – ni leurs interconnexions ne sont figés. Les *data papers* n'échappent pas à cette règle ; de fait, on observe différents cas de figure en ce qui concerne leur publication. Sur le plan éditorial, les *data papers* sont publiés soit dans des revues de données (*data journals*) dédiées majoritairement à ce type d'articles, soit dans des revues non spécifiquement dédiées aux *data papers*, dont le type est, selon le cas, plus ou moins bien identifié c'est-à-dire avec ou sans rubrique *data papers*. Le contenu du *data paper* peut également varier. Une partie des *data papers* peut, par exemple, s'avérer complexe, voire proche des articles de recherche, avec des parties qui exposent les résultats d'analyse, une discussion, un état de l'art, etc., alors que d'autres *data papers* sont courts et visent essentiellement à restituer les métadonnées du *template* proposé par la revue. Par ailleurs, il existe des types d'articles qu'on pourrait situer entre les articles de recherche et les *data papers*, comme des *data services papers* (des articles sur les services de données) ou des *overlay articles*, publiés en partenariat avec des entrepôts de données, qui peuvent parfois être confondus avec les *data papers* à proprement parlé.

3. En ce qui concerne leur diffusion, la plupart des *data papers* sont publiés et diffusés sur des plateformes de revues, comme c'est le cas pour les articles de recherche. Certaines plateformes de revues commencent à proposer des services de stockage et de publication des données signalées par un *data paper* et/ou par un article de recherche, et au moins une plateforme publie également des plans de gestion (*data management plans*) et teste depuis plusieurs années l'import de *data papers* générés plus ou moins automatiquement à partir des métadonnées d'entrepôts de données (Pamerlon dans ce livre).

4. L'interconnexion et la convergence entre plateformes de revues et entrepôts de données sont des facteurs centraux de la dynamique de cet écosystème toutes disciplines confondues. Plusieurs plateformes ont développé des partenariats forts, voire même une intégration verticale avec des entrepôts de données. D'autres revues recommandent de déposer les données sur des entrepôts « fiables », certifiés, dignes de confiance (Marongiu *et al.* dans ce livre), sécurisés et/ou opérés par des organismes, universités ou instituts de recherche de réputation (Schöpfel et Rebouillat, 2023).

5. On notera en effet que deux autres types de documents contiennent des éléments d'information fortement structurée et proches, voire similaires à l'information contenue dans un *data paper*. Il s'agit d'une part, du plan de gestion des données (*data management plan*, DMP) dont notamment la version finale, au terme du projet de recherche, contient des informations détaillées sur la production et le type des données, sur leur traitement et sur leur conservation, informations qui pourraient être utilement récupérées lors de la rédaction d'un *data paper* sur les mêmes données (León y Barella dans ce livre). D'autre part, les dossiers de recherche à destination des comités d'éthique pour l'évaluation de la conformité d'un projet scientifique avec les principes éthiques contiennent également des informations sur la nature et le traitement des données qui présentent un intérêt certain pour le *data paper*. On pourrait donc imaginer dans l'avenir une interconnexion entre ces trois types de documents.

Un paysage diversifié

Article de recherche et *data paper*

6. La variante la plus courante et la plus proche du *data paper* est l'article traditionnel de recherche accompagné des données dont il traite. Dans ce cas, les données peuvent être hébergées par la revue en tant que matériel supplémentaire ou déposées dans un entrepôt tiers (Kratz et Strasser, 2015). Parfois, ces matériaux supplémentaires à l'article (*supplementary material*), faute d'espace dédié sur le serveur des revues en ligne, se retrouvent quelque part sur d'autres sites, maintenus par exemple par les auteurs. Le projet *NISO/NFAIS Supplemental Journal Article Materials* visait à développer des recommandations pour l'inclusion, la manipulation, l'affichage et la préservation par les éditeurs des matériaux supplémentaires (Kunze *et al.*, 2011). Mais la simple citation des données dans l'article traditionnel ne permet pas forcément aux lecteurs d'obtenir toutes les informations sur la nature et l'utilité des données, sur les normes et standards, etc. (Callaghan, 2013). En effet, les métadonnées des données mentionnées dans les articles classiques ne suffisent pas pour une réutilisation, d'où l'importance du *data paper* (Candela *et al.*, 2015).

Publication par procuration

7. Les articles qui contiennent une description générale d'une base de données ou d'un jeu de données, avec ou sans analyse, ont également été désignés comme des « publications de données par procuration » (*data publication by proxy*) ; par procuration (*by proxy*) veut dire ici que ces articles fournissent une référence bibliographique à citer (avec un DOI), mais ne contiennent pas les données directement (Lawrence *et al.*, 2011 ; Mooney, 2016).

8. Une autre variante dans le domaine du traitement automatique des langues, du nom de « feuille de données » (*datasheet for dataset*), est également proche des *data papers*, dans la mesure où elle contient plusieurs de leurs caractéristiques, notamment la motivation (l'objectif) de la création des données, composition et recommandations (confidentialité par exemple), les modalités de la collecte et du traitement, l'accessibilité, les limites et contraintes (Gebru *et al.*, 2021).

9. D'autres types d'articles se distinguent de la définition des *data papers*, comme les articles sur les services de données (*data services papers*) qui concernent davantage les entrepôts de données et d'autres services de stockage.

Les articles exécutables

10. La crise de la reproductibilité alliée à la prise en main de plus en plus fréquente des langages informatiques, tel que Python, par les chercheurs en SHS, a créé un nouveau besoin d'où découle l'apparition d'une autre variante du *data paper* : l'article exécutable (*computational paper*) qui semble être la variante la plus en vogue en 2022. La possibilité d'intégrer du calcul dans un récit au sein d'un document électronique suscite un certain engouement chez les chercheurs, notamment en SHS. En effet, le code informatique est lisible par le lecteur, jouable et réplicable. Ces articles, pour la plupart adossés à des cahiers électroniques qui, dans le même document, peuvent rassembler du texte, des images, des formules mathématiques et du code informatique exécutable (*notebook jupyter*), consistent en un code rédigé, accompagné de toutes les données d'entrée nécessaires, alterné avec des morceaux de texte le commentant (Arnaud *et al.*, dans ce livre). L'intérêt de ce type de publication est triple, elle permet de rendre l'expérience vérifiable et reproductible immédiatement, elle permet de mettre en situation les données et de comprendre les suites de raisonnement scientifique et elle permet enfin au lecteur d'avoir réuni en une seule publication le commentaire et les données. Cependant, l'intérêt de publier les deux types d'information, scientifique et technique, ensemble et au même endroit fait l'objet de débat, car il ne s'agit pas nécessairement du même type de lecteur qui cherche ces informations (Callaghan, 2013).

11. Pour répondre à cette problématique, le *Journal of Digital History*[1] a développé un format de publication qui se situe entre le *data paper*, l'article traditionnel et l'article exécutable. L'objectif est double : aller plus loin qu'un simple *data paper* et proposer une réplicabilité totale en rendant les résultats reproductibles grâce au système des *notebooks jupyter* et ouvrir les boîtes noires de la recherche dans toutes les étapes du processus (Clavert et Fickers, 2022). Ainsi, la revue propose différentes couches (*layers*) qui peuvent être masquées ou non, selon le besoin et l'intérêt du lecteur.

12. Une dernière question se pose quant aux articles exécutables, à savoir la mise à jour et le *versioning*. En effet, ces articles utilisent des librairies à un instant T qui rendent la lecture future plus complexe, nécessitant des ajustements ou des changements en fonction des packages utilisés. Or, la mise à jour de ces publications avec différentes versions ne correspond pas aux pratiques habituelles des revues scientifiques qui elles-mêmes publient à un temps T. La possibilité d'une interconnexion évolutive entre le *data paper* exécutable et les librairies conduirait nécessairement à réformer les modalités et le fonctionnement des publications, en même temps qu'elle générerait de nouvelles questions juridiques, comme l'adaptation du droit d'auteur aux différentes versions.

1. https://journalofdigitalhistory.org/en/articles

Les expositions de données

13. L'écosystème des *data papers* ne serait pas complet si nous omettions de mentionner un dernier type de document qui accompagne parfois le *data paper* dans un souci de médiation. Il semble d'ailleurs que c'est à travers ce document que les principes FAIR[2] se matérialisent le mieux. Il s'agit, plus précisément, des pages web indépendantes du *data paper*, mais hébergées sur le même serveur que la revue. La revue existe toujours mais tout le reste – data exhibit et ScieMedia – semble avoir disparu. À remplacer par : Ce phénomène se retrouve principalement au sein de la communauté de chercheurs en computer vision ou NLP à travers les Project Pages ou « page de présentation du projet ». Il s'agit de pages web indépendantes offrant une brève présentation d'un jeu de données, agrémentées de multimédias intégrés, d'éléments interactifs et de fonctionnalités permettant de prévisualiser et d'explorer directement le jeu de données décrit[3]. Les auteurs proposent à travers ce média une nouvelle manière de présenter leur travail et permettent à des utilisateurs de saisir en un coup d'œil l'objet du travail de recherche sans avoir à télécharger intégralement le jeu de données ou lire l'article complet. On peut considérer qu'il s'agit là d'une version allégée d'un executable paper puisque seuls quelques visualisations ou quelques morceaux de code sont mis à disposition.

2. Principes FAIR cf. https://www.go-fair.org/.
3. https://detection-based-text-line-recognition.github.io/ ou https://antoyang.github.io/vidchapters.html.

14. Ces Project Pages n'ont pas de patterns identiques et varient grandement d'un chercheur à l'autre en fonction de son institution de rattachement. Elles sont ainsi rarement hébergées sur le même serveur que l'article de données présenté et le sont le plus souvent sur GitHub. Une expérimentation pour les SHS sur un modèle similaire à l'informatique, a été conduite au sein de la revue Research Data Journal for the Humanities and Social Sciences à travers une plateforme d'expositions (showcase) de données[4]. Cette expérience a duré de 2016 à 2020. Les auteurs de celle-ci justifient son arrêt par la difficulté au sein des SHS de prise en compte de fichiers supplémentaires à l'article, qui doit être limité à des cas occasionnels (Breure, Leen, Peter Doorm et Hans Voorbij, 2022). Les éditeurs comme Brill et Pensoft se sont intéressés à la manière de présenter des articles de données mais la question de l'archivage de ces pages web restent problématiques.ent les données, le *data paper* expliquant l'ensemble. On trouve aussi, le cas échéant, un site web pour découvrir davantage le projet dans le cadre duquel ont été produites les données. Selon le cas, un lien permet d'explorer les données. On peut considérer qu'il s'agit là d'une version allégée d'un *executable paper* puisque seuls quelques visualisations ou quelques morceaux de code sont mis à disposition avec pour avantage d'explorer une partie des données tout en saisissant très vite leur intérêt sans avoir toutefois à installer le logiciel associé.

4. https://ssh.datastations.nl/dataset.xhtml?persistentId=doi:10.17026/SS/TLTMIR

L'automatisation du *data paper*

15. L'Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE) a utilisé pour la mise en place de son entrepôt de données *Data INRAE* l'application en open source Dataverse. Il a développé à cette occasion une fonctionnalité supplémentaire qui permet la génération automatique d'un *data paper* à partir des métadonnées des dépôts dans un format texte ouvert propre à INRAE ou compatible avec la revue *Data in Brief*. L'entrepôt Recherche Data Gouv lancé à l'été 2022 qui s'appuie sur le Dataverse d'INRAE a également repris cette fonctionnalité (Kosmopoulos et Schöpfel dans ce livre).

16. L'automatisation de la production d'un *data paper* est donc une réalité, au moins partiellement. Quels sont les enjeux ? Tout d'abord techniquement, la génération d'un *data paper* exige un degré élevé de normalisation et d'interopérabilité entre les entrepôts de données, les outils de traitement de texte et les plateformes de revues. Certes, un outil pour faciliter la conversion d'un document avec des métadonnées en un manuscrit d'article serait avantageux pour certaines disciplines, par exemple pour les sources en biologie (Pamerlon, dans ce livre). Toutefois, scientifiquement, il semble compliqué d'accepter la soumission de *data papers* produits automatiquement par machines pour certaines revues (Cottineau-Mugadza *et al.*, dans ce livre) en suivant le concept du *data author* proposé par *The New England Journal of Medicine* (Bierer *et al.*, 2017).

17. Comme le montrent les différentes contributions à cet ouvrage, en SHS en particulier, la valeur ajoutée humaine est significative et généralement indispensable, pour décrire notamment la valeur des données et leur (re) utilisation potentielle. Les informations contextuelles sont importantes et la granularité des informations dépend des projets et de leurs porteurs. Or, certains entrepôts exigent des métadonnées très riches, d'autres se contentent d'un minimum standard, compatible avec le format de DataCite[5], par exemple.

18. Actuellement, on peut observer la mise en place d'une sorte de chaîne de valeur des données de recherche où les *data papers* ont un rôle spécifique à jouer : certaines métadonnées font partie des fichiers de données, tandis que d'autres sont ajoutées par l'entrepôt au moment du dépôt ou plus tard ; d'autres informations, plus riches et d'origine diverse (plan de gestion, dossier éthique, article de recherche…), sont ajoutées dans les *data papers* et par les revues de données qui contribuent ainsi à ce que les données soient plus faciles à trouver, à accéder et à réutiliser, par une information contextuelle (*paradata*) complémentaire aux métadonnées dans les entrepôts (Huvila, 2022).

19. Il est trop tôt pour dire si (et comment) cette chaîne de valeur se stabilisera dans la durée. Automatiser au moins partiellement la production d'un *data paper* pourrait favoriser une certaine standardisation, comme c'est déjà le cas pour un article de recherche. Dans ce cas, l'article généré automatiquement pourrait servir de squelette normé au futur *data paper*. Toutefois, l'automatisation seule n'apporte pas de réponse scientifiquement satisfaisante et ne répond pas non plus à l'enjeu majeur des différences disciplinaires dans la gestion et la publication des données.

Métadonnées et FAIRisation

20. La gestion des métadonnées qui décrivent le jeu de données joue un rôle fondamental, puisque celles-ci ouvrent la voie à la réutilisation des données par des chercheurs même extérieurs à la discipline. L'enjeu de cette gestion se joue à deux niveaux :

- Standardisation : La qualité des *data papers* dépend en grande partie de la qualité des métadonnées des données, c'est-à-dire des terminologies contrôlées (nomenclatures, etc.), des formats standard, des éléments (champs) bien définis. On peut aussi supposer que le développement des *data papers* et des revues de données s'accompagnera de réflexions sur les normes des métadonnées des données, sur les articles et les revues, tout autant par les chercheurs, que les éditeurs et les professionnels de l'information.
- Spécialisation : Les normes relatives aux métadonnées doivent également être aussi conformes que possible aux exigences et caractéristiques spécifiques des diverses pratiques, disciplines, méthodes, outils et équipements

5. https://schema.datacite.org/

scientifiques. Cette spécialisation tend à limiter leur interopérabilité entre différents domaines, infrastructures, systèmes d'information. Certains entrepôts adoptent une approche qu'on pourrait qualifier comme « aussi spécifique que possible, aussi générique que nécessaire ». Une autre approche consiste à accepter deux (ou plusieurs) normes différentes pour chaque jeu de données et chaque *data paper*, l'une générique (notamment, le format DataCite), l'autre spécifique.

21. Dans les deux cas, standardisation et spécialisation, l'application des principes FAIR reste centrale. Ils permettent de comprendre le rôle joué par les *data papers* pour la réutilisation potentielle des données. L'objectif principal des principes FAIR est l'interconnexion des infrastructures de données, notamment dans le cadre de l'European open science Cloud (Kosmopoulos, 2022). La standardisation des métadonnées est au cœur des enjeux et projets de FAIRisation (Rivet dans ce livre). Si les principes FAIR concernent en premier lieu les entrepôts de données, les plateformes de publication des revues de données ou de revues qui publient des *data papers* sont également concernées. En effet, les principes FAIR ont pour objectif de faciliter non seulement l'accès aux données, mais également l'interconnexion entre les articles de recherche, les données et les *data papers*.

22. Les *data papers* apportent un surcroît d'informations par rapport aux métadonnées de l'entrepôt et contribuent à la réalisation des principes FAIR notamment de deux :

- La facilité de trouver les données (*findability*) : les métadonnées des données sont décrites d'une manière riche (principe F2), et elles sont enregistrées et indexées dans une ressource interrogeable (principe F4).
- La facilité de réutiliser les données (*reusability*) : les métadonnées sont richement décrites avec une pluralité d'attributs précis et pertinents (principe R1), c'est-à-dire qu'elles sont publiées avec une licence d'utilisation des données claires et accessibles (principe R1.1), et elles sont associées à une provenance détaillée (principe R1.2).

23. Pour mémoire, faciliter la réutilisation des données est l'objectif primaire d'un *data paper*. On comprend d'autant plus le rôle crucial de la FAIRisation des métadonnées pour le développement des *data papers*. Plus indirectement, les *data papers* contribuent également à l'accessibilité des données (principe A2 : les métadonnées resteront accessibles sur une plateforme de revues même si les données ne le sont plus) et à l'interopérabilité des données, dans la mesure où les métadonnées utilisent un langage normalisé et contiennent des liens vers d'autres métadonnées et données (principes I1 et I3). On pourrait imaginer dans l'avenir une interconnexion plus étroite entre la structure du *data paper* avec ses différentes composantes et les modèles (*templates*) proposés par les entrepôts de données de manière à faciliter l'alimentation de l'un par l'autre, et cela en adoptant différentes variables qui prennent en compte le niveau de spécialisation décrit plus haut.

Le cas des SHS

24. Si la question de nouveaux modes de publication de données touche l'ensemble des sciences (cf. Kaden, 2019), les SHS apparaissent en retrait par rapport à la gestion des données de recherche (Schöpfel *et al.*, 2022). Les enjeux scientifiques et économiques de l'ouverture des données concernent davantage les sciences de la vie et la médecine. On peut se demander si l'incitation à l'ouverture des données n'est pas prématurée par rapport aux besoins des chercheurs dont l'évaluation et la reconnaissance restent toujours principalement fondées sur la publication d'articles et d'ouvrages, et peut-être les services dédiés aux données sont-ils encore relativement éloignés des communautés de recherche (Rebouillat, 2019). À ceci s'ajoutent en SHS deux particularités : l'importance de la dimension interprétative pour l'encodage des données, pour les métadonnées descriptives et pour les annotations d'une part, et l'importance de la contextualisation des données en question, c'est-à-dire la connaissance de leur acquisition, production ou construction, de la contribution des personnes à l'origine des données, etc., d'autre part (Schöpfel, 2020). C'est justement ce dernier aspect qui rend les *data papers* malgré tout attractifs pour les SHS et pourrait leur permettre de faire un saut qualitatif dans le processus de validation scientifique. En effet, l'absence d'un « savoir contextuel » sur les données, leur création, leur traitement et leur utilisation est un obstacle majeur pour leur réutilisation, en particulier pour les SHS qui utilisent une grande variété de données.

25. Les chercheurs et les chercheuses sont toutefois encore peu nombreux à publier des *data papers*, y compris en France, où les organismes de recherche, les SCD et les Urfist, commencent à proposer des formations[6]. Leur développement dépendra sans doute plutôt des incitations au partage des données avec une vraie reconnaissance institutionnelle, de l'évolution des moyens de partage, comme les entrepôts de données, des moteurs de recherche, des outils de citations, etc., que des revues scientifiques. En d'autres termes, si les chercheurs déposent davantage leurs données dans les entrepôts, cela devrait les inciter à publier davantage de *data papers*. Les « méga-revues de données » comme *Data in Brief* offrent de leur côté une solution rapide et efficace ; mais d'autres projets d'édition sont en cours de montage, plus appropriés sans doute à la demande des SHS en France, comme notamment la revue *Données, Expériences, Méthodes et Codes* (*DEMC*) sur la plateforme *Numerev*[7], ou la revue interdisciplinaire Data & Corpus – La revue des données en SHS, portée par l'Université de Lorraine et accessible sur la plateforme Episciences[8].

26. Au moment où les SHS se trouvent à un tournant quantitatif face à l'accroissement potentiel des sources et de nouveaux moyens de recherche grâce au traitement de texte et aux algorithmes, les *data papers* peuvent contribuer à mieux (faire) comprendre et exposer les méthodes et les résultats de la recherche en SHS. Favoriser

6. Voir le chapitre d'Alicia León Y Barella dans cet ouvrage.
7. https://demc-journal.org/
8. https://msh-lorraine.fr/mshl-17/

l'écriture d'articles de données au sein des revues SHS contribuerait à consolider les disciplines concernées tout en les valorisant à travers des normes interopérables et des procédures de standardisation. Jacques Revel dans *Épistémologie des Sciences sociales* conclut son article par ces mots : « Tout récit est, en ce sens, une opération de modélisation. Cette opération est le plus souvent tacite. Elle aurait beaucoup à gagner à ne pas le rester », et il incite les scientifiques des SHS à rendre compréhensibles et structurées des méthodes disciplinaires qui ne l'étaient pas toujours (Revel, 2012). Écrire un *data paper* en SHS, c'est rendre visibles les opérations faites par les chercheurs et affirmer la scientificité des faits établis.

27. On peut supposer qu'avec le déploiement de la science ouverte et l'évolution des critères d'évaluation des chercheurs et chercheuses selon la déclaration de San Francisco (DORA), le nombre de *data papers* devrait continuer d'augmenter. Les enjeux à venir seront surtout du côté de la standardisation des formats, de la convergence et de l'interconnexion des entrepôts et des plateformes de publication, du développement de nouvelles fonctionnalités par l'anonymisation pour la partie évaluation, mais également de la terminologie et des identifiants, de la gouvernance et de l'ouverture des infrastructures. Les risques déjà perceptibles sont toutefois une fragmentation du paysage avec une privatisation croissante des systèmes et des outils, et une publication prédatrice des données et des *data papers* favorisée par le modèle économique de l'accès ouvert avec APC.

Liste des abréviations/ acronymes

1. TRF-GIS : Third-Republic France Geographic information System (Système d'Information Géographique de la France de la Troisième République)
2. FAIR : Findable, Accessible, Interoperable, Reusable
3. ANR : Agence Nationale de la Recherche
4. SIG : Système d'Information Géographique
5. PUD : Plateforme Universitaire de Données
6. CNRS : Centre National de la Recherche Scientifique
7. URFIST : Unité régionale de formation à l'information scientifique et technique
8. INRAE : Institut national de recherche pour l'agriculture, l'alimentation et l'environnement
9. MSHS : Maison des Sciences de l'Homme et de la Société
10. CHSP : Centre d'histoire de Sciences Po
11. API - Interface de programmation d'application
12. CC - Creative Commons
13. ADPIC - Accord sur les aspects des droits de propriété intellectuelle qui touchent au commerce
14. CC0 - Creative Commons Zéro
15. CCPL - Licence publique Creative Commons
16. CC REL - Langage d'expression des droits Creative Commons
17. CE - Commission européenne
18. CJUE - Cour de justice de l'Union européenne
19. DDO - La Directive sur les données ouvertes
20. DPI - Droits de propriété intellectuelle
21. DRM - Gestion des droits numériques
22. Droit *sui generis* - Droits *sui generis* sur les bases de données
23. FAIR - Traçabilité, accessibilité, interopérabilité et réutilisabilité
24. FLOSS - Free/Libre Open Source Software
25. GDT - Exploration de textes et de données

26. GNU GPL - GNU General Public License
27. IoT - Internet of Things/Internet des objets
28. MP3 - MPEG-1 Audio Layer 3
29. MTP - Mesures techniques de protection
30. OMPI - Organisation Mondiale de la Propriété Intellectuelle
31. RGPD - Règlement général sur la protection des données
32. CLARIN : Common Language Resources and Technology Infrastructure
33. SSHOC : Social Sciences and Humanities Open Cloud
34. LIBER : Ligue des Bibliothèques Européennes de Recherche
35. SQL : Structured Query Language

Bibliographie

La bibliographie est accessible au format Zotero par ce lien : https://nakala.fr/10.34847/nkl.6ebf0l2r.

Adamczak, Beata. 2021. « Data Journals and Data Papers in Various Research Areas and Scientific Disciplines – Bibliometric Analysis Based on InCites ». *TASK Quaterly* 25 (4) : 433– 471. https://doi.org/10.34808/tq2021/25.4/p.

Agostini-Marchese, Enrico, Emmanuel Château-Dutier et Michael Sinatra. 2021. « Nouvelles perspectives sur les humanités numériques ». 2021. Sens Public. http://sens-public.org/dossiers/1596/.

Akers, Katherine. 2014. « A growing list of data journals ». *Data@MLibrary* (blog). https://mlibrarydata.wordpress.com/2014/05/09/data-journals.

Akrich, Madeleine. 1992. « The De-scription of Technical Objects ». Dans *Shaping Technology/Building Society. Studies in Sociotechnical Change*, édité par Wiebe E. Bijker et John Law Bijker, 205-224. Cambridge : MIT Press. https://halshs.archives-ouvertes.fr/halshs-00081744.

Ali, Nawel Aït et Jean-Pierre Rouch. 2013. « Le "je suis débordé" de l'enseignant-chercheur. Petite mécanique des pressions et ajustements temporels ». *Temporalités* 18. https://doi.org/10.4000/temporalites.2632.

Allés-Torrent, Susan, Gimena del Rio Riande, Jerry Bonnel, Dieyun Song et Nidia Hernández. 2021. « Digital Narratives of COVID-19: A Twitter Dataset for Text Analysis in Spanish ». *Journal of Open Humanities Data* 7 : 5. https://doi.org/10.5334/johd.28.

Almas, Bridget et Thibault Clérice. 2017. « Continuous Integration and Unit Testing of Digital Editions ». *Digital Humanities Quarterly* 11 (4).

Analyzed Layout and Text Object (ALTO) (v4.2). 2020. https://www.loc.gov/standards/alto/news.html#4-2-released.

André, Virginie. 2019. « Pourquoi faire de la sociolinguistique des interactions verbales avec des enseignants et des apprenants de Français Langue Étrangère ? ». *Linx. Revue des linguistes de l'université Paris X Nanterre* 79. https://doi.org/10.4000/linx.3694.

Association Francophone des Utilisateurs de Logiciels Libres. « Groupe de travail sur l'interopérabilité ». 2015. https://aful.org/gdt/interop.

Association des archivistes français, section Aurore. 2016. « Référentiel de gestion des archives de la recherche ». https://www.archivistes.org/Referentiel-de-gestion-des-archives-de-la-recherche.

Aston, Guy et Lou Burnard. 2020. The BNC Handbook : Exploring the British National Corpus with SARA. Édimbourg : Edinburgh University Press. https://doi.org/10.1515/9780748628889.

Atici, Levent, Sarah Whitcher Kansa, Justin Lev-Tov et Eric C. Kansa. 2013. « Other People's Data : A Demonstration of the Imperative of Publishing Primary Data ». *Journal of Archaeological Method and Theory* 20 (4) : 663–681. https://doi.org/10.1007/s10816-012-9132-9.

Baker, Monya. 2016. « 1 500 Scientists Lift the Lid on Reproducibility ». *Nature* 533 : 452-454. https://doi.org/10.1038/5334452a.

Bakker, Ryan, Catherine de Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen et Milada Anna Vachudova. 2015. « Measuring Party Positions in Europe : The Chapel Hill Expert Survey Trend File, 1999-2010 ». *Party Politics* 21 (1) : 143–152. https://doi.org/10.1177/1354068812462931.

Baloup, Julie, Emre Bayamlıoğlu, Aliki Benmayor, Charlotte Ducuing, Lidia Dutkiewicz, Teodora Lalova, Yuliya Miadzvetskaya et Bert Peeters. 2021. « White Paper on the Data Governance Act ». CiTiP Working Paper, 38. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3872703.

Barba, Lorena A. 2018. « Terminologies for Reproducible Research ». *ArXiv*. http://arxiv.org/abs/1802.03311.

Barone, Carlo. 2021. « Relative Risk Aversion Models : How Plausible are their Assumptions ? Review of Top-Cited Articles ». https://doi.org/10.21410/7E4/UA0UKM, *data.sciencespo*, V1.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, *et al.* 2018. « Reproducible Research in Linguistics : A Position Statement on Data Citation and Attribution in Our Field ». *Linguistics* 56 (1) : 1-18. https://doi.org/10.1515/ling-2017-0032.

Berne Convention for the Protection of Literary and Artistic Works (adopted 14 July 1967, entered into force 29 January 1970) 828 UNTS 221 (Berne Convention)

Bierer, Barbara E., Mercè Crosas et Heather H. Pierce. 2017. « Data Authorship as an Incentive to Data Sharing ». *New England Journal of Medicine* 376 (17) : 1684-1687. https://doi.org/10.1056/NEJMsb1616595.

Boillet, Mélodie, Marie-Laurence Bonhomme, Dominique Stutzmann et Christopher Kermorvant. 2019. « HORAE : an annotated dataset of books of hours ». Communication présentée à *The 5th International Workshop on Historical Document Imaging and Processing*, 7-12. Sydney, Australia : ACM Press. https://doi.org/10.1145/3352631.3352633.

Bordelon, Dominic, Uta Grothkopf, Silvia Meakins et Michael Sterzik. 2016. « Trends and developments in VLT data papers as seen through telbib ». Dans *Observatory Operations : Strategies, Processes, and Systems VI*, édité par Alison B. Peck, Robert L. Seaman et Chris R. Benn, 9910 : 811-817. SPIE. https://doi.org/10.1117/12.2231697.

Borgman, Christine L. 2015. *Big data, little data, no data. Scholarship in the networked world*. Cambridge MA : MIT Press.

Borgman, Christine L. 2012. « The conundrum of sharing research data ». *Journal of the American Society for Information Science and Technology* 63 (6) : 1059-1078. https://doi.org/10.1002/asi.22634.

Bornmann, Lutz. 2014. « Do Altmetrics Point to the Broader Impact of Research? An Overview of Benefits and Disadvantages of Altmetrics ». *Journal of Informetrics* 8 (4) : 895-903. https://doi.org/10.1016/j.joi.2014.09.005.

Boukacem-Zeghmouri, Chérifa et Françoise Paquienséguy. 2021. « Les Humanités numériques, depuis les Sic ». Dans *Questionner les Humanités numériques : Positions et propositions des SIC*, édité par Nicolas Pélissier et Françoise Paquienséguy, 57-72. SFSIC et CPdirsic. https://www.sfsic.org/wp-inside/uploads/2021/06/questionner-humanites-numeriques.pdf.

Bracco, Laetitia. 2022. « Mesurer l'ouverture de la science : le cas de l'Université de Lorraine ». *Revue française des sciences de l'information et de la communication* 24. https://doi.org/10.4000/rfsic.12474

Brazil, « Law n. 9.610 of February 19, 1998 (Law on Copyright and Neighboring Rights) ».

Breure, Leen, Peter Doorn et Hans Voorbij. 2021. « Data Showcases : the Data Journal in aMultimodal World ». *International Journal of Digital Curation* 17 (1). http://dx.doi.org/10.2218/ijdc.v17i1.789.

Brouard, Sylvain, Martial Foucault, Elie Michel. 2020. « Citizens' Attitudes Under COVID-19 Pandemic ». https://doi.org/10.21410/7E4/EATFBW, data.sciencespo, V13, UNF:6:K8+KwbSpnlfyAhoOH8YHDQ== [fileUNF]

Burnard, Lou. 2002. « Where did we Go Wrong? A Retrospective Look at the British National Corpus ». Dans *Teaching and Learning by Doing Corpus Analysis*, édité par Bernhard Ketteman and Georg Marko, 51-70. Amsterdam : Rodopi. https://doi.org/10.1163/9789004334236_007.

Callaghan, Sarah. 2013. « Data Journals - as Soon-to-Be-Obsolete Stepping Stone to Something Better ? ». *Citing Bytes - Adventures in Data Citation* (blog). http://citingbytes.blogspot.com/2013/01/data-journals-as-soon-to-be-obsolete.html.

Callaghan, Sarah, Steve Donegan, Sam Pepler, Mark Thorley, Nathan Cunningham, Peter Kirsch, Linda Ault, *et al*. 2012. « Making Data a First-Class Scientific Output : Data Citation and Publication by NERC's Environmental Data Centres ». *International Journal of Digital Curation* 7 (1) : 107–113. https://doi.org/10.2218/ijdc.v7i1.218.

Candela, Leonardo, Donatelli Castelli, Paolo Manghi et Alice Tani. 2015. « Data Journals : A Survey ». *Journal of the Association for Information Science and Technology* 66 (9) : 1747-1762. https://doi.org/10.1002/asi.23358.

Carroll, Michael W. 2007. « Creative Commons as Conversational Copyright ». Dans *Intellectual Property and Information Wealth : Issues and Practices in the Digital Age* édité par Peter K. Yu (ed), 445-461. Praeger. Villanova Law/Public Policy Research Paper No. 2007-8. https://ssrn.com/abstract =978813.

Carroll, Michael W. 2015. « Sharing Research Data and Intellectual Property Law : A Primer ». PLoS Biology 13 (8) e1002235. https://doi.org/10.1371/journal.pbio.1002235.

Case C-338/02 Fixtures Marketing Ltd v Svenska Spel AB [2004] ECR I-10497.

Case C-203/02 The British Horseracing Board Ltd and Others v William Hill Organization Ltd [2004] ECR I-10415 .

Catherine, Hugo. 2023. « Étude comparative des services nationaux de recherche ». Dans *Partage et valorisation des données de la recherche. Développements, tendances et modèles*, édité par Joachim Schöpfel et Violaine Rebouillat, p. 147-166. Londres, ISTE.

Chagué, Alix, Victoria Le Fourner, Manuela Martini et Éric Villemonte de la Clergerie. 2019. « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non uniforme ? ». Communication présentée à *DH-Nord 2019 Corpus et archives numériques*. Lille. https://hal.inria.fr/hal-02448921.

Chagué, Alix, Lucas Terriel et Laurent Romary. 2020. « Des images au texte : LECTAUREP, un projet de reconnaissance automatique d'écriture ». Poster présenté à *DH-nord 2020 The Measurement of Images. Computational Approaches in the History and Theory of the Arts*. Lille. https://hal.archives-ouvertes.fr/hal-03008579.

Champaud, Christian. 1994. « The Development of Verb Forms in French Children at around Two Years of Age: Some Comparisons with Romance and Non-Romance

Languages ». Communication présentée au *First Lisbon Meeting on Child Language*, Lisbon, Portugal.In.

Chang, Andrew C. et Phillip Li. 2022. « Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not." ». *Critical Finance Review* 11 (1) : 185–206. https://doi.org/10.1561/104.00000053.

Chavan, Vishwas S. et Peter Ingwersen. 2009. « Towards a data publishing framework for primary biodiversity data : challenges and potentials for the biodiversity informatics community ». *BMC Bioinformatics* 10, 14 : S2. https://doi.org/10.1186/1471-2105-10-S14-S2.

Chavan, Vishwas et Lyubomir Penev. 2011. « The Data Paper : a Mechanism to Incentivize Data Publishing in Biodiversity Science ». *BMC Bioinformatics*, 12 (Suppl. 15) : S2. https://doi.org/10.1186/1471-2105-12-S15-S2.

Chen, Tianqi et Carlos Guestrin. 2016. « XGBoost : A Scalable Tree Boosting System ». Dans Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. San Francisco : Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Chen, Ya-Ning. 2017. « An analysis of characteristics and structures embedded in data papers ». *Libellarium* 9 (2) : 145–156. https://doi.org/10.15291/libellarium.v9i2.266.

Chin, George et Carina S. Lansing. 2004. « Capturing and Supporting Contexts for Scientific Data Sharing via the Biological Sciences Collaboratory ». *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 409–418. https://doi.org/10.1145/1031607.1031677.

Christensen, Garret et Edward Miguel. 2018. « Transparency, Reproducibility, and the Credibility of Economics Research ». *Journal of Economic Literature* 56 (3) : 920-980. https://doi.org/10.1257/jel.20171350.

Cioni, Martina, Giovanni Federico et Michelangelo Vasta. 2020. « The Long-Term Evolution of Economic History : Evidence from the Top Five Field Journals (1927-2017) ». *Cliometrica* 14 (1) : 1-39. https://doi.org/10.1007/s11698-019-00186-x.

Clavert, Frederic et Andreas Fickers. 2022. « Publishing digital history scholarship in the era of updatism ». *Journal of Digital History* 2 (1). https://doi.org/10.1515/JDH-2022-0003.

Clavert, Frédéric et Valérie Schafer. 2019. « Les humanités numériques, un enjeu historique ». *Quaderni* 98 (1) : 33-49. https://doi.org/10.4000/quaderni.1417.

Clérice, Thibault et Alix Chagué. 2021. *HUM Generator, the HTR United Metadata Generator* (v0.0.1). Python. https://doi.org/10.5281/zenodo.5363307.

Clérice, Thibault et Ariane Pinche. 2021a. *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects* (v0.0.4). Python. https://doi.org/10.5281/zenodo.5356154.

Clérice, Thibault et Ariane Pinche. 2021b. *HTRVX, HTR Validation with XSD* (version 0.0.1). Python. https://doi.org/10.5281/zenodo.5359963.

CNRS. 2018. Une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche. Étude COPIST 2. Rapport, CNRS DIST, Paris

Collins, S. *et al*. European Commission. European Commission Expert Group on FAIR Data. 2018. « Turning FAIR into reality : Final Report and Action Plan from the European Commission Expert Group on FAIR Data ». https://

op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/

Comité pour la science ouverte. « Rapport du Groupe de Travail sur les Cahiers de Laboratoire électroniques ». 2021. https://www.ouvrirlascience.fr/rapport-du-groupe-de-travail-sur-les-cahiers-de-laboratoire-electroniques/.

Coriat, Benjamin. 2011. « From Natural-Resource Commons to Knowledge Commons : Common Traits and Differences ». *LEM Papers Series* 16 : 22. https://ideas.repec.org/p/ssa/lemwps/2011-16.html.

CoSO. 2019. « Pour un politique des données de la recherche : guide stratégique ». *Ouvrir la science* (blog). https://www.ouvrirlascience.fr/pour-une-politique-des-donnees-de-la-recherche-guide-strategique-a-lusage-des-etablissements.

CoSO. 2021. *Étude de faisabilité d'un service générique d'accueil et de diffusion des données*. Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, Comité pour la science ouverte, Paris.

Costello, Mark J., William K. Michener, Mark Gahegan, M., Zhi-Qiang Zhang et Philip E. Bourne. 2013. « Biodiversity Data Should Be Published, Cited, and Peer Reviewed ». *Trends in Ecology & Evolution*, 28 (8) : 454–461. https://doi.org/10.1016/j.tree.2013.05.002.

Creative Commons. 2019. « About CC Licenses ». https://creativecommons.org/about/cclicenses/.

Creative Commons. 2022. « CC Attribution 4.0. International ». https://creativecommons.org/licenses/by/4.0/legalcode.

Creative Commons. 2022a. « Use & Remix ». https://creativecommons.org/use-remix/.

Creative Commons. 2022b. « What does it mean that Creative Commons Licenses are "machine-readable"? ». https://creativecommons.org/faq/#what-does-it-mean-that-creative-commons-licenses-are-machine-readable.

Creative Commons. 2022c. « 'License Chooser' ». https://creativecommons.org/choose/.

Creative Commons. 2022d. « Public Domain Mark 1.0. ». https://creativecommons.org/publicdomain/mark/1.0/.

Creative Commons. 2022e. « CC0 1.0. Universal ». https://creativecommons.org/publicdomain/zero/1.0/legalcode.

Creative Commons Wiki. 2021. « CC REL ». https://wiki.creativecommons.org/wiki/CC_REL.

Dacos, Marin *et al.* 2010. « Manifeste des *Digital humanities* ». *THATCamp Paris* (blog). https://doi.org/10.58079/u027.

Dacos, Marin et Pierre Mounier. 2010. « Les carnets de recherche en ligne, espace d'une conversation scientifique décentrée ». Dans *Lieux de savoir, T.2, Gestes et supports du travail savant*. Paris : Albin Michel.

Dalbera, Jean-Phillipe. 2002. « Le corpus entre données, analyse et théorie ». *Corpus* 1. https://doi.org/10.4000/corpus.10

Dani Arribas-Bel, Seraphim Alvanides, Michael Batty, Andrew Crooks, Linda See et Levi Wolf. 2021. « Urban Data/Code : A new EP-B Section ». *Environment and Planning B : Urban Analytics and City Science* 48 (9) : 2517-2519. https://doi.org/10.1177/23998083211059670.

Data.gouv.fr. 2021. « Printemps de data.gouv.fr : nos réflexions sur la qualité des données ». https://www.data.gouv.fr/fr/posts/nos-reflexions-sur-la-qualite-des-donnees/.

Deboin, Marie-Claude. 2018. « Se familiariser avec les plans de gestion de données de la recherche ». CIRAD. https://doi.org/10.18167/coopist/0056.

Dedieu, Laurence. 2014. *Rédiger et publier un data paper dans une revue scientifique en 5 points*. Montpellier (FRA) : CIRAD, 7 p. https://collaboratif.cirad.fr/alfresco/s/d/workspace/SpacesStore/75735aae-dacb-4052-833c-9017a2b2bba4/rediger-et-publier-un-data-paper-octobre2014.pdf.

Dedieu, Laurence. 2017. « Revues publiant des *data papers* ». Actualités – Coopérer En Information Scientifique et Technique – Cirad. https://coop-ist.cirad.fr/contenus-annexes/documents/revues-publiant-des-data-papers-nov-2017.

Dehousse, Renaud, Selma Bendjaballah, Geneviève Michaud, Olivier Rozenberg, Florence Deloche-Gaudez, Giuseppe Ciavarini Azzi, Olivier Costa et Romain Lalande. 2017. « L'*Observatory of European Institutions* : une base de données sur le processus décisionnel dans l'Union européenne (1996-2014) ». *Politique européenne* 58 (4) : 14-42. https//doi.org/10.3917/poeu.058.0014.

Desrosières, Alain. 2010 [1993]. *La politique des grands nombres : Histoire de la raison statistique*. Paris : La Découverte. https://doi.org/10.3917/dec.desro.2010.01.

Dietrich, Karin et Marie-Héléne Varnier. 1987. « Les Allemands naturalisés en France de 1791-1848. Méthodologie et résultats statistiques ». *Cahiers d'études germaniques* 13 : 9-56. https://doi.org/10.3406/cetge.1987.1020

DIM Matériaux anciens et patrimoniaux – PPSM (CNRS, ENS, Paris-Saclay). 2021. *Projets soutenus/CREMMA*. DIM MAP. https://www.dim-map.fr/projets-soutenus/cremma/.

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996], OJ L 77/20 (Database Directive)

Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, OJ L 157/1 (Trade Secret Directive).

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019], OJ L 130/92 (CDSM).

Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast) [2019], OJ L 172/56.

Dosi, Giovanni et Joseph Stiglitz. 2013. « The Role of Intellectual Property Rights in the Development Process, with Some Lessons from Developed Countries : An Introduction ». *LEM Working Paper Series* 23. https://www.econstor.eu/bitstream/10419/89516/1/771928769.pdf.

Drahos, Peter. 2016. *A Philosophy of Intellectual Property*. Hants : Dartmouth Publishing Company.

Drexl, Joseph, Reto Hilty, Jure Globocnik, *et al.* 2017. « Position Statement of the Max Planck Institute for Innovation and Competition of 26 April 2017 on the European Commission's "Public consultation on Building the European Data Economy ». *Max Planck Institute for Innovation & Competition Research Paper No. 17-08*. https://ssrn.com/abstract =2959924.

Drexl, Joseph, Reto Hilty, Francisco Beneke, Luc Desaunettes-Barbero, *et al.* 2019. « Technical Aspects of Arti-

ficial Intelligence : An Understanding from an Intellectual Property Law Perspective ». *Max Planck Institute for Innovation and Competition Research Paper Series No. 19-13*. https://ssrn.com/abstract =3465577.

Drucker, Johanna. 2011. « Humanities Approaches to Graphical Display ». *Digital Humanities Quarterly* 5 (1) : 1-21.

Duchesne, Sophie, Guillaume Garcia, Anne Both et Sarah Cadorel. 2014. « Retour vers le futur : la numérisation des enquêtes qualitatives de sciences sociales entre patrimonialisation et transformation des pratiques scientifiques ». *Huma-Num, Le blog d'Huma-Num et des consortiums-HN* (blog). https://humanum.hypotheses.org/147.

Ducuing C., T, Margoni, L. Schirru, D. Spajic, T. Lalova-Spinks, L. Stähler, E. Bayamlıoğlu, A. Pétel, J. Chu, B. Peeters, A. Christofi, J. Baloup, M. Avramidou, A. Benmayor, T. Gils, E. Kun, E. De Noyette et E. Biasin. 2022. « White Paper on the Data Act Proposal ». https://www.law.kuleuven.be/citip/en/news/item/old/white-paper-data-act.

Dusollier, Séverine. 2017. « Propriété Inclusive ou Inclusivité ». Dans *Dictionnaire des biens communs*, édité par Marie Cornu, Fabienne Orsi et Judith Rochfeld, 983-987. Paris : Presses Universitaires de France..

Dusollier, Séverine. 2006. « The Master's Tools v. The Master's House : Creative Commons v. Copyright ». *Columbia Journal of Law & Arts*, 29. https://papers.ssrn.com/sol3/papers.cfm?abstract_id =2186187.

Earp, Brian D. et David Trafimow. 2015. « Replication, Falsification, and the Crisis of Confidence in Social Psychology ». *Frontiers in Psychology* 6. https://doi.org/10.3389/fpsyg.2015.00621.

Egloff, Willi, Donat Agosti, David Patterson, Anke Hoffmann, Daniel Mietchen, Puneet Kishor et Lyubomir Penev. 2016. « Data Policy Recommendations for Biodiversity Data. EU BON Project Report ». *Research Ideas and Outcomes 2*, e8458. https://doi.org/10.3897/rio.2.e8458.

Ekama, Kate, Johan Fourie, Hans Heese et Lisa-Cheree Martin. 2021. « When Cape Slavery Ended : Introducing a New Slave Emancipation Dataset ». *Explorations in Economic History* 81 : 101390. https://doi.org/10.1016/j.eeh.2021.101390.

Engelhardt, Claudia, Katarzyna Biernacka, Aoife Coffey, Ronald Cornet, Alina Danciu, Yuri Demchenko, Stephen Downes, *et al.* 2022. *D7.4 How to be FAIR with your Data. A Teaching and Training Handbook for Higher Education Institutions* (version V1.2 DRAFT). Zenodo. https://doi.org/10.5281/zenodo.5905866.

Établissements de recherche et d'enseignement supérieur français. 2015. « Charte française de déontologie des métiers de la recherche ». https://comite-ethique.cnrs.fr/charte/.

European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act)' COM/2022/68 final ('Data Act proposal').

Faniel, Ixchel M., Rebecca D. Frank et Elizabeth Yakel. 2019. « Context From the Data Reuser's Point of View ». *Journal of Documentation*, 75 (6) : 1274-1297. https://doi.org/10.1108/JD-08-2018-0133.

Farace, Dominic J. et Joachim Schöpfel. 2020. Data Papers provide an Innovative Tool for Information and Data Management. Poster. *GL 2020 Twenty-Second International Conference on Grey Literature "Applications of Grey*

Literature for Science and Society", Nov. 2020, Rome, Italy. https://hal.science/hal-03825967v1.

Ferger, Anne et Hanna Hedeland. 2020. « Towards Continuous Quality Control for Spoken Language Corpora ». *International Journal of Digital Curation* 15 (1). https://doi.org/10.2218/ijdc.v15i1.601.

Fisher, William. 1987. « Theories of Intellectual Property ». https://cyber.harvard.edu/people/tfisher/iptheory.pdf.

Frank, Rebecca D., Yakel, Elizabeth et Ixchel M. Faniel. 2015. « Destruction/reconstruction : Preservation of Archaeological and Zoological Research Data ». *Archival Science* 15 (2) : 141-167. https://doi.org/10.1007/s10502-014-9238-9.

Friedhoff, Stefan, Christian Meier zu Verl, Christian Pietsch, Christian Meyer, Johanna Vompras et Stefan Liebig. 2013. *Social Research Data : Documentation, Management, and Technical Implementation within the SFB 882* vol. 16. SFB 882 Working Paper Series. Bielefeld : DFG Research Center (SFB) 882 From Heterogeneities to Inequalities. https://pub.uni-bielefeld.de/record/2560035 Gray, Stephen. 2015. *Case Study : Publishing a Data Paper*. DOCZZ. http://doczz.net/doc/6729472/case-study%E2%80%94publishing-a-data-paper---research-data-service Jefferies, Neil, Fiona Murphy, Anusha Ranganathan et Hollydawn Murray. 2019. « Data2paper : Giving Researchers Credit for Their Data ». *Publications* 7 (2) : 36. https://doi.org/10.3390/publications7020036

Furet, François. 1971. « L'histoire quantitative et construction du fait historique ». *Annales. Histoire, Sciences Sociales* 26 (1) : 63-75. https://doi.org/10.3406/ahess.1971.422459.

Gabay, Simon, Jean-Baptiste Camps, Ariane Pinche et Claire Jahan. 2021. « SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more) ». Communication présentée au *1st International Workshop on Computational Paleography (IWCP@ICDAR 2021)*. Lausanne, Switzerland. https://hal.archives-ouvertes.fr/hal-03336528.

Gabay, Simon, Thibault Clérice et Christian Reul. 2020. *OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more)*. Preprint. https://hal.archives-ouvertes.fr/hal-02577236.

Gaillard, Jeanne. 1997. *Paris, la ville (1852-1870)*. Paris : L'Harmattan.

Gandrud, Christopher. 2015. *Reproducible Research with R and Rstudio*. 2e éd. Boca Raton : CRC Press.

Garcia-Garcia, Alicia, Alexandre López-Borrul et Fernanda Peset. 2015. « Data journals : eclosión de nuevas revistas especializadas en datos ». *Profesional de la información* 24 (6) : 845-854. https://doi.org/10.3145/epi.2015.nov.17.

Gärtner Anette et Kate Brimsted. 2017. « Let's Talk about Data Ownership ». *European Intellectual Property Review* 39 (8) : 461-466.

Gaudiaut, Tristan. 2021. « Le Big Bang du Big Data ». *Statistica*. https://fr.statista.com/infographie/17800/big-data-evolution-volume-donnees-numeriques-genere-dans-le-monde/.

Gay, Victor. 2021. « Mapping the Third Republic : A Geographic Information System of France (1870-1940) ». *Historical Methods : A Journal of Quantitative and Interdisciplinary History* 54 (4) : 189-207. https://doi.org/10.1080/01615440.2021.1937421.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III et Kate Crawford. 2021. « Datasheets for datasets ».

Communications of the ACM 64 (12) : 86–92. https://doi.org/10.1145/3458723.

Github Inc. 2021. *Fork a Repo*. Documentation. GitHub Docs. https://docs.github.com/en/get-started/quickstart/fork-a-repo.

GNU Operating System. 2022. « GPL-Compatible Free Software Licenses ». https://www.gnu.org/licenses/license-list.html#GPLCompatibleLicenses.

Gozlan, Clémentine. 2016. « Les sciences humaines et sociales face aux standards d'évaluation de la qualité académique. Enquête sur les pratiques de jugement dans une agence française ». *Sociologie* 7 (3) : 261-280. https://doi.org/10.3917/socio.073.0261.

Graber-Soudry, Ohad, Timo Minssen, Daniel Nilsson, Marcelo Corrales. 2021. « Legal Interoperability and the FAIR Data Principles (1.0) ». https://doi.org/10.5281/zenodo.4471312.

Grandini, Margherita, Enrico Bagli et Giorgio Visani. 2020. « Metrics for Multi-Class Classification : An Overview ». *ArXiv*. http://arxiv.org/abs/2008.05756.

Grandjonc, Jacques. 1972. « Éléments statistiques pour une étude de l'immigration ». *Archiv für Sozialgeschichte*, 12 : 487-533.

Grandjonc, Jacques. 1974. « Les étrangers à Paris sous la monarchie de Juillet et la seconde République ». In *Population* 29(1) : 61-88.

Grandjonc, Jacques. 1983. *Émigrés français en Allemagne, Émigrés allemands en France 1685-1945*. Paris : Institut Goethe.

Granger, Sabrina. 2020. « Au-delà de la reproductibilité : la transparence de la recherche ». Forge. GitLab Inria Learning Lab Mooc Recherche Reproductible. https://gitlab.inria.fr/learninglab/mooc-rr/mooc-rr-ressources/blob/master/module0/ressources/ResearchTransparency_fr.org.

Green, Nancy. 1998. « Du Sentier à la 7e Avenue : la confection et les immigrés ». Paris – New York.

Habert, Benoit. 2010. « Construire ensemble des mémoires numériques durables : l'archivage numérique pérenne ». *Dans 13e Congrès International sur le Document Électronique (CIDE)*, édité par Madjid Ihadjadene, Manuel Zacklad, Khaldoun ZREIK, 5-24. Paris : Europia Productions. https://halshs.archives-ouvertes.fr/halshs-00991508.

Hadrossek, Christine, Joana Janik, Maurice Libes, Violaine Louvet, Marie-Claude Quidoz, Alain Rivet et Geneviève Romier. 2021. « Guide de bonnes pratiques sur la gestion des données de la recherche ». HAL Id : hal– 03152732. https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html.

Hall-Lew, Lauren, Claire Cowie, Stephen McNulty, Nina Markl, Shan-Jan Sarah Liu, Catherine Lai, ClaireLlewellyn, *et al.* 2021. « The Lothian Diary Project : Investigating the Impact of the COVID-19 Pandemic on Edinburgh and Lothian Residents ». *Journal of Open Humanities Data* 7 : 4. https://doi.org/10.5334/johd.25.

Hey, Tony, Stewart Tansley et Kristin Tolle, éd. 2009. *The fourth paradigm. Data-intensive scientific discovery*. Redmond, WA : Microsoft Corp. http://research.microsoft.com/en-us/collaboration/fourthparadigm/ Huang, Pao Pei et Wei Jeng. 2022. « Data Paper as a Reward ? Motivation, Consideration, and Perspective behind Data Paper Submission». *Proceedings of the Association for Information Science and Technology* 59 (1) : 437-441. https://doi.org/10.1002/pra2.648.

Hooghe, Lisbet, Ryan Bakker, Anna Brigevich, Catherine de Vries, Erica Edwards, Gary Marks, Jan Rovny, Marco Steenbergen, Milada Vachudova. 2010. « Reliability and Validity of the 2002 and 2006 Chapel Hill Expert Surveys on Party Positioning ». *European Journal of Political Research* 49 (5) : 687-703. https://doi.org/10.1111/j.1475-6765.2009.01912.x.

Houot, Isabelle et Emmanuel Triby. 2020. *Le chercheur en activités*. Collection Didactique, humanités et sciences pour l'ingénieur. Université de technologie de Belfort-Montbéliard. https://hal.univ-lorraine.fr/hal-03158548.

Hsu, Leslie, Raleigh L. Martin, Brandon McElroy, Kimberly Litwin-Miller et Wonsuck Kim. 2015. « Data Management, Sharing, and Reuse in Experimental Geomorphology : Challenges, Strategies, and Scientific Opportunities. *Geomorphology* 244 : 180-189. https://doi.org/10.1016/j.geomorph.2015.03.039.

HTR-United. (2020) 2022. *HTR-United Catalog*. YAML. https://github.com/HTR-United/htr-united/blob/026e680323b47f6206a6d6007cb96d6cc756fab5/htr-united.yml.

HTR-United, Alix Chagué et Thibault Clérice. 2021. *HTR-United: Ground Truth Resources for the HTR of patrimonial documents*. https://github.com/HTR-United/htr-united.

Hugenholtz, P. Bernt. 2018. « Against "data property ». Dans *Kritika : Essays on Intellectual Property, Volume 3*, édité par Hanns Ullrich, Peter Drahos et Gustavo Ghidini. Edward Elgar Publishing. https://doi.org/10.4337/9781788971164.00010

Huggett, Jeremy. 2018. « Reuse Remix Recycle : Repurposing Archaeological Digital Data ». *Advances in Archaeological Practice* 6 (2) : 93-104. https://doi.org/ 10.1017/aap.2018.1.

Hutson, Scott R. 2006. « Self-Citation in Archaeology : Age, Gender, Prestige, and the Self ». *Journal of Archaeological Method and Theory* 13 (1) : 1-18. https://doi.org/10.1007/s10816-006-9001-5.

Huvila, Isto. 2016. « Awkwardness of becoming a boundary object: Mangle and materialities of reports, documentation data, and the archaeological work ». *The Information Society* 32 (4) : 280-297. https://doi.org/10.1080/01972243.2016.1177763.

Huvila, Isto. 2022. « Improving the Usefulness of Research Data with Better Paradata ». *Open Information Science* 6(1) : 28-48. https://doi.org/10.1515/opis-2022-0129.

Hwang, Lorraine, Allison Fish, Laura Soito, MacKenzie Smith et Louise H. Kellogg. 2017. « Software and the Scientist : Coding and Citation Practices in Geodynamics ». *Earth and Space Science* 4 (11) : 670-680. https://doi.org/10.1002/2016EA000225.

Ichter, Jean, Olivier Gargominy, Marie-France Leccia, Solène Robert et Laurent Poncet. 2022. « The First Large-Scale All Taxa Biodiversity Inventory in Europe : Description of the Mercantour National Park ATBI Datasets ». *Biodiversity Data Journal* 10 : e85901. https://doi.org/10.3897/BDJ.10.e85901.

Ide, Nancy. 2013. « An Open Linguistic Infrastructure for Annotated Corpora ». Dans *The People's Web Meets NLP. Theory and Applications of Natural Language Processing* édité par Julia Hirschberg, Eduard Hovy, Mark Johnson, 265–285. Berlin, Heidelberg : Springer. https://doi.org/10.1007/978-3-642-35085-6_10.

Ioannidis, John P. A. 2005. « Why Most Published Research Findings Are False ». *PLoS Medicine* 2 (8) : e124. https://doi.org/10.1371/journal.pmed.0020124.

Jahan, Claire et Simon Gabay. 2021. *OCR17 +* (v1.0). https://github.com/e-ditiones/OCR17plus.

Jiao, Chenyue et Peter T. Darch. 2020. « The Role of the Data Paper in Scholarly Communication ». *Proceedings of the Association for Information Science and Technology* 57 (1) : e316. https://doi.org/10.1002/pra2.316.

Joachims, Thorsten. 1998. « Text Categorization with Support Vector Machines : Learning with many Relevant Features ». Dans *Machine Learning : ECML-98*, édité par Claire Nédellec et Céline Rouveirol, 137-142. Berlin : Springer.

Jurafsky, Daniel et James H. Martin. 2008. Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition with Language Models. 2e éd. Upper Saddle River : Pearson. https://web.stanford.edu/~jurafsky/slp3/.

Kaden, Ben. 2019. « Pourquoi les données de recherche ne sont-elles pas publiées ? ». *Études de communication* 52, 137-146. https://doi.org/10.4000/edc.8783.

Kahle, Philip, Sebastian Colutto, Günter Hackl, et Günter Mühlberger. 2017. « Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents ». Communication présentée à *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 19-24. https://doi.org/10.1109/ICDAR.2017.307.

Kai, Li et Chenyue Jiao. 2021. « The Data Paper as a Sociolinguistic Epistemic Object : A Content Analysis on the Rhetorical Moves Used in Data Paper Abstracts". *Journal of the Association for Information Science and Technology* 73 (6) : 834-846. https://doi.org/10.1002/asi.24585.

Kansa, Sarah Whitcher et Eric C. Kansa. 2018. « Data Beyond the Archive in Digital Archaeology: An Introduction to the Special Section ». *Advances in Archaeological Practice* 6 (2) : 89-92. https://doi.org/10.1017/aap.2018.7.

Karila-Cohen, Karine, Claire Lemercier, Isabelle Rosé et Claire Zalc. 2018. « Nouvelles Cuisines de l'histoire Quantitative ». *Annales. Histoire, Sciences Sociales* 73 (4) : 773-783. https://doi.org/10.1017/ahss.2019.90.

Kembellec, Gérald. 2019. « Semantic publishing, la sémantique dans la sémiotique des codes sources d'écrits d'écran scientifiques ». *Les Enjeux de l'information et de la communication* 20 (2) : 55-72. https://doi.org/10.3917/enic.027.0055

Kembellec, Gérald et Olivier Le Deuff. 2022. « Poétique et ingénierie des *data papers* ». *Revue française des sciences de l'information et de la communication* 24. https://doi.org/10.4000/rfsic.12938.

Kembellec, Gérald et Olivier Le Deuff. 2022. « Data Paper : Émergence d'Une Nouvelle Donne Scientifique ». *Revue Française des sciences de l'information et de la communication* 24. https://doi.org/10.4000/rfsic.12219.

Kembellec, Gérald et Olivier Le Deuff. 2022. « Poétique et ingénierie des *data papers* ». *Revue française des sciences de l'information et de la communication* 24. https://doi.org/10.4000/rfsic.12938.

Kerber, Wolfgang. 2016. « A New (Intellectual) Property Right for Non-Personal Data ? An Economic Analysis ». *MAGKS Joint Discussion Paper Series in Economics* 37-2016.

Khan, Nushrat, Mike Thelwall et Kayvan Kousha. 2021. « Measuring the Impact of Biodiversity Datasets : Data

Reuse, Citations and Altmetrics ». *Scientometrics* 126 : 3621–3639. https://doi.org/10.1007/s11192-021-03890-6.

Kiessling, Benjamin. (2015) 2021. *mittagessen/kraken: 3.0.5* (v3.0.5). Python. https://github.com/mittagessen/kraken.

Kim, Jihyun. 2020. « An Analysis of Data Paper Templates and Guidelines : Types of Contextual Information Described by Data Journals ». *Science Editing* 7 (1) : 16–23. https://doi.org/10.6087/kcse.185.

Kim, Jihyun, Elizabeth Yakel et Ixchel M. Faniel. 2019. « Exposing Standardization and Consistency Issues in Repository Metadata Requirements for Data Deposition ». *College & Research Libraries* 80 (6) : 843-875. https://doi.org/10.5860/crl.80.6.843

Knauf, Audrey. 2022. Appréhender l'écriture d'un Data Paper proposant une approche réflexive et analytique des données de la recherche en SHS. *Webinaire Data papers : quand ? comment ? pourquoi ?, GTSO Couperin*. https://hal.science/hal-03773080

Knöchelmann, Marcel. 2019. « Open science in the Humanities, or: Open Humanities? » *Publications* 7 (4): 65. https://doi.org/10.3390/publications7040065.

Knuth, D. E. 1984. « Literate programming ». *The Computer Journal* 27 (2) : 97-111. https://doi.org/10.1093/comjnl/27.2.97.

Knuutila, Aleksi, Aliaksandr Herasimenko, Hubert Au, Jonathan Bright et Philip N. Howard. 2021. « A Dataset of COVID-Related Misinformation Videos and their Spread on Social Media ». *Journal of Open Humanities Data* 7: 6. http://doi.org/10.5334/johd.24.

Kobrock, Kristina et Timo B. Roettger. 2022. « Assessing the Replication Landscape in Experimental Linguistics ». *PsyArXiv*. https://doi.org/10.31234/osf.io/fzngs.

König, Mareike. 2003. *Deutsche Handwerker, Arbeiter und Dienstmädchen in Paris*. Munich : Oldenbourg Wissenschaftsverlag. https://doi.org/10.1524/9783486834383.

König, Mareike. 2006. « Les Allemands à Paris au XIXe siècle ». *Annuaires de l'École pratique des hautes études* 20 : 387-389. https://www.persee.fr/doc/ephe_0000-0001_2004_num_20_1_11528.

Kosmopoulos Christine. 2022. « From Open Access Publishing to open science. An Overview of the Last Developments in Europe and in France ». Dans *Handbook of Research on the Global View of Open Access and Scholarly Communications* édité par Daniel Gelaw Alemneh, 1-22, IGI Global https://doi.org/10.4018/978-1-7998-9805-4.ch001.

Kosmopoulos, Christine et Michèle Dassa, éd. 2011. Actes du colloque « Évaluation des productions scientifiques. Des innovations en SHS ? ». https://shs.hal.science/halshs-03556892v1.

Kotti, Zoe, Konstantinos Kravvaritis, Konstantina Dritsa et Diomidis Spinellis. 2020. « Standing on shoulders or feet ? An extended study on the usage of the MSR data papers ». *Empirical Software Engineering* 25 (5) : 3288– 3322. https://doi.org/10.1007/s10664-020-09834-7.

Kotti, Zoe et Diomidis Spinellis. 2019. « Standing on Shoulders or Feet? The Usage of the MSR Data Papers ». Dans *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 565-576. Montreal, QC, Canada. https://doi.org/10.1109/MSR.2019.00085.

Kratz, John et Carly Strasser. 2014. « Data Publication Consensus and Controversies ». *F1000Research* 3 : 94. https://doi.org/10.12688/f1000research.3979.3.

Kratz, John Ernest et Carly Strasser. 2015. « Researcher Perspectives on Publication and Peer Review of Data ». *PLOS ONE* 10 (2) : e0117619. https://doi.org/10.1371/journal.pone.0117619.

Kronauge, F.-A. 1855. *Adreßbuch der Deutschen in Paris für das Jahr 1854 oder vollständiges Adreßverzeichniß aller in Paris und seinen Vorständten wohnenden selbständigen Deutschen in alphabetischer Ordnung*. Paris : Selbstverlag des Verfassers. https://bibliotheques-specialisees.paris.fr/ark :/ 73873/pf0000884072.

Kučera, Henry, W. Nelson Francis, W. F. Twaddell, Mary Lois Marckworth, Laura M. Bell et John B. Carroll, and. 1967. *Computational Analysis of Present Day American English*. 1St Edition. Providence, RI : Brown University Press.

Kunze, John A, Patricia Cruse, Rachael Hu, Stephen Abrams, Kirt Hastings, Catherine Mitchell et Lisa R. Schiff. 2011. *Practices, Trends, and Recommendations in Technical Appendix Usage for Selected Data-Intensive Disciplines*. UC Office of the President : California Digital Library. https://escholarship.org/uc/item/9jw4964t

Laine, Christine, Steven N. Goodman, Michael E. Griswold et Harold C. Sox. 2007. « Reproducible Research : Moving Toward Research the Public can Really Trust ». *Annals of Internal Medicine* 146 (6) : 450-453. https://doi.org/10.7326/0003-4819-146-6-200703200-00154.

Landi, Annalisa, Mark Thompson, Viviana Giannuzzi, Fedele Bonifazi, Ignasi Labastida, Luiz Olavo Bonino da Silva Santos, Marco Roos. 2020. « The "A" of FAIR – As open as possible, as closed as necessary ». *Data Intelligence* 2 (1-2) : 47-55.

LARHRA. 2011. *Les cantons français de 1884 à 1966*. Lyon : Laboratoire de Recherche Historique Rhône-Alpes. http://geo-larhra.ish-lyon.cnrs.fr/?q =atlas-historique/territoires-d-etat/evolution-des-cantons-en-france.

Lasser, Jana. 2020. « Creating an Executable Paper is a Journey through open science ». *Communications Physics* 3 (1) : 1-5. https://doi.org/10.1038/s42005-020-00403-4.

Lawrence, Bryan, Catherine Jones, Brian Matthews, Sarah Pepler et Sarah Callaghan. 2011. « Citation and Peer Review of Data : Moving Towards Formal Data Publication ». *International Journal of Digital Curation* 6 (2) : 4-37. https://doi.org/10.2218/ijdc.v6i2.205.

Le Bail, Hélène et Calogero Giametta. 2021. « Enquête inter-associative sur l'impact de la Loi prostitution de 2016 : entretiens associations (1/3) ». https://data.sciencespo.fr/dataverse/elp2016, data.sciencespo, V2.

Le Deuff, Olivier. 2018. « Une nouvelle rubrique pour la RFSIC : le *data paper* ». *Revue française des sciences de l'information et de la communication* 15. http://journals.openedition.org/rfsic/5275.

Lee, Jungyeoun et Jihyun Kim. 2021. « Korean Researchers' Motivations for Publishing in Data Journals and the Usefulness of their Data : a Qualitative Study. *Science Editing* 8 (2) : 145–152. https://doi.org/10.6087/kcse.246.

Leeuwen, Marco H. D., Ineke Maas, Andrew Miles et Sören Edvinsson. 2002. *HISCO : Historical International Standard Classification of Occupations*. Louvain : Leuven University Press.

Le Fourner, Victoria, Joachim Schöpfel. 2025. « Le paysage des *data papers* ». Dans *Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, édité par Christine Kosmopoulos et Joachim Schöpfel. Villeneuve-d'Ascq : Presses universitaires du Septentrion. https://www.septentrion.com/fr/livre/?GCOI=27574100316700.

Légifrance. 2021. « Décret n° 2021-1572 du 3 décembre 2021 relatif au respect des exigences de l'intégrité scientifique par les établissements publics contribuant au service public de la recherche et les fondations reconnues d'utilité publique ayant pour activité principale la recherche publique ». https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044411360.

Lemercier, Claire et Claire Zalc. 2008. *Méthodes quantitatives pour l'historien*. Paris : La Découverte (Repères). https://doi.org/10.3917/dec.lemer.2008.01.

Lemercier, Claire et Claire Zalc. 2019. *Quantitative Methods in the Humanities : An Introduction*. Traduit par Arthur Goldhammer. Charlottesville : University of Virginia Press. https://doi.org/10.2307/j.ctvbqs963.

Lepareur, Fanny, Mathieu Manceau, Yorick Reyjol, Julien Touroult, Solène Robert, Frédéric Vest, Arnaud Horellou et Laurent Poncet, L. 2022. The Nationwide 'ZNIEFF' Inventory in France : an Open Dataset of More than One Million Species data in Zones of High Ecological Value. *Biodiversity Data Journal* 10 : e71222. https://doi.org/10.3897/BDJ.10.e71222.

L'Hostis, Dominique, Marjolaine Hamelin, Virginie Lelievre et Pascal Aventurier. 2016. *Publier un Data Paper pour valoriser ses données*. Support de formation Infodoc Express. https://doi.org/10.15454/1.478247389988942E12.

Li, Kai, Jane Greenberg et Jillian Dunic. 2020. « Data Objects and Documenting Scientific Processes : An Analysis of Data Events in Biodiversity Data Papers». *Journal of the Association for Information Science and Technology* 71 (2) : 172–182. https://doi.org/10.1002/asi.24226.

Li, Kai, Jane Greenberg et Xia Lin. 2016. « Software Citation, Reuse and Metadata Considerations : an Exploratory Study Examining LAMMPS ». Dans *Proceedings of the Association for Information and Technology* 53 (1), 1-10. Silver Spring, Maryland : American Society for Information Science. https://doi.org/10.1002/pra2.2016.14505301072.

Li, Kai et Chenyue Jiao. 2022. « The Data Paper as a Sociolinguistic Epistemic Object : A Content Analysis on the Rhetorical Moves used in Data Paper Abstracts ». *Journal of the Association for Information Science and Technology* 73 (6) : 834–846. https://doi.org/10.1002/asi.24585.

Li, Kai, Chuyi Lu et Chenyue Jiao. 2021. « A Survey of Exclusively Data Journals and How They Are Indexed by Scientific Databases». *Proceedings of the Association for Information Science and Technology* 58 (1) : 771–773. https://doi.org/10.1002/pra2.557.

Li, Kai et Erjia Yan. 2018. « Co-Mention Network of R Packages : Scientific Impact and Clustering Structure ». *Journal of Informetrics* 12 (1) : 87-100. https://doi.org/10.1016/j.joi.2017.12.001.

Lüdecke, Daniel, Alexander Bartel, Carsten Schwemmer, Chuck Powell, Amir Djalovski et Johannes Titz. 2022. « sjPlot : Data Visualization for Statistics in Social Science ». https://cran.r-project.org/web/packages/sjPlot/index.html.

MacWhinney, Brian et Catherine Snow. 1985. « The Child Language Data Exchange System ». *Journal of*

Child Language 12 (2) : 271-295. https://doi.org/10.1017/S0305000900006449.

MacWhinney Brian et Catherine Snow. 1990. « The Child Language Data Exchange System ». *ICAME Journal* 14 : 3-25.

Maignien, Yannick. 2011. « ISIDORE, de l'interconnexion de données à l'intégration de services ». https://archivesic.ccsd.cnrs.fr/sic_00593320.

Maniadis, Zacharias et Fabio Tufano. 2017. « The Research Reproducibility Crisis and Economics of Science ». *The Economic Journal* 127 (605) : F200-F208. https://doi.org/10.1111/ecoj.12526.

Maniadis, Zacharias, Fabio Tufano et John A. List. 2017. « To Replicate or Not to Replicate ? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study ». *The Economic Journal* 127 (605) : F209-F235. https://doi.org/10.1111/ecoj.12527.

Margoni, Thomas. 2016. « The Harmonisation of EU Copyright Law : The Originality Standard ». Dans *Global Governance of Intellectual Property in the 21st Century*, 85-105. Springer International Publishing. https://doi.org/10.1007/978-3-319-31177-7_6.

Margoni, Thomas, Diane Peters. 2016. « Creative Commons Licenses : Empowering Open Access ». *SSRN*. https://dx.doi.org/10.2139/ssrn.2746044.

Margoni, Thomas, Martin Kretschmer. 2022. « A Deeper Look into the EU Text and Data Mining Exceptions : Harmonisation, Data Ownership, and the Future of Technology ». *GRUR International* 71(8) : 685-701. https://doi.org/10.1093/grurint/ikac054.

Margoni, Thomas et Mark Perry. 2010. « Free-Libre Open Source Software as a Public Policy Choice ». *International Journal on Advances in Internet Technology* 3 (3,4) : 212-222.

Mariotti, Viola. 2020. « Transcription automatique des feuillets du Manuscrit du Roi ». Billet. *ANR Maritem* (blog). 19 octobre 2020. https://maritem.hypotheses.org/193.

Marongiu, Paola, Nilo Pedrazzini, Marton Ribary, et Barbara McGillivray. 2023. « Le Journal of Open Humanities Data (JOHD) : Enjeux Et Défis Dans La Publication De Data Papers Pour Les Sciences Humaines Et Sociales (SHS) ». Zenodo. https://doi.org/10.5281/zenodo.7624854.

Massot, Marie-Laure, Arianna Sforzini et Vincent Ventresque. 2019. « Transcribing Foucault's handwriting with Transkribus ». *Journal of Data Mining and Digital Humanities*, Atelier Digit_Hum. https://hal.archives-ouvertes.fr/hal-01913435.

Mayernik, Matthew S., Sarah Callaghan, Roland Leigh, Jonathan Tedds et Steven Worley. 2015. « Peer Review of Datasets : When, Why, and How ». *Bulletin of the American Meteorological Society* 96 (2) : 191–201. https://doi.org/10.1175/BAMS-D-13-00083.1.

McGillivray, Barbara, Paola Marongiu, Nilo Pedrazzini, Marton Ribary, Mandy Wigdorowitz et Eleonora Zordan. 2022. « Deep Impact : A Study on the Impact of Data Papers and Datasets in the Humanities and Social Sciences ». *Publications* 10 (4) : 39. https://doi.org/10.3390/publications10040039.

Mejias, Ulises A. et Nick Couldry. 2019. « Datafication ». *Internet Policy Review* 8 (4). https://doi.org/10.14763/2019.4.1428

MESRI. 2021. *Deuxième Plan national pour la science ouverte*. Paris : ministère de l'Enseignement supérieur, de la Re-

cherche et de l'Innovation. https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte.

Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. 2021. « Deuxième plan national pour la science ouverte (2021-2024) ». 2021. https://www.ouvrirlascience.fr/wp-content/uploads/2021/06/Deuxieme-Plan-National-Science-Ouverte_2021-2024.pdf.

Moch, Leslie Page. 2008. « Frankreich ». In *Enzyklopädie Migration in Europa. Vom 17. Jahrhundert bis zur Gegenwart*, édité par Klaus Bade, 122-41.

Mooney, Hailey. 2016. « Scholarly Communication and Data ». Dans *Databrarianship : The academic data librarian in theory and practice* édité par Lynda Kellam et Kristi Thompson, 195–218. Chicago : Association of College and Research Libraries.

Mpondo-Dicka, Patrick. 2020. « Le *Markdown*, une praxis énonciative du numérique ». *Interfaces numériques* 8 (2) : 304-304. https://doi.org/10.25965/interfaces-numeriques.3915.

Neuroth, Heike, Stefan Strathmann, Achim Oßwald et Jens Ludwig, éd. 2013. *Digital Curation of Research Data. Experiences of a Baseline Study in Germany*. Glückstadt : Verlag Werner Hülsbusch. http://www.nestor.sub.uni-goettingen.de/bestandsaufnahme/Digital_Curation.pdf.

Noiriel, Gérard. 1992. *Population, immigration et identité nationale en France : XIX^e^-XX^e^ siècle*. Paris : Hachette.

Noizet, Hélène ; Bove, Boris et Costa, Laurent. 2013. *Paris de parcelles en pixels. Analyse géomatique de l'espace parisien médiéval et moderne*, Paris : Presses universitaires de Vincennes, Comité d'histoire de la Ville de Paris.

Obels, Pepijn, Daniël Lakens, Nicolas A. Coles, Jaroslav Gottfried, Seth A. Green. 2020. « Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology ». *Advances in Methods and Practices in Psychological Science* 3(2) : 229-237. https://doi.org/10.1177/2515245920918872.

OECD. 2007. *OECD Principles and Guidelines for Access to Research Data from Public Funding*, Paris : OECD Publishing. https://doi.org/10.1787/9789264034020-en-fr.

OCDE. 1996. *L'économie fondée sur le savoir*. Paris.

Open Science Collaboration. 2015. « Estimating the Reproducibility of Psychological Science ». *Science* 349 (6251) : aac 4716. https://doi.org/10.1126/science.aac4716.

Organisation de coopération et de développement économiques. 2007. Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. Paris : Éditions OCDE. https://doi.org/10.1787/9789264034020-en-fr.

Ostrom, Elinor. 1999. « Private and Common Property Rights ». https://www.sfu.ca/~allen/common%20property.pdf.

Pan, Xuelian, Erjia Yan, Qianqian Wang et Weina Hua. 2015. « Assessing the Impact of Software on Science: A Bootstrapped Learning of Software Entities in Full-Text Papers ». *Journal of Informetrics* 9 (4) : 860-871. https://doi.org/10.1016/j.joi.2015.07.012.

Park, Hyoungjoo et Dietmar Wolfram. 2019. « Research Software Citation in the Data Citation Index: Current Practices and Implications for Research Software Sharing and Reuse ». *Journal of Informetrics* 13 (2) : 574-582. https://doi.org/10.1016/j.joi.2019.03.005.

Park, Hyoungjoo, Sukjin You et Dietmar Wolfram. 2018. « Informal Data Citation for Data Sharing and Reuse is More Common than Formal Data Citation in Biomedical Fields ». *Journal of the Association for Information Science and Technology* 69 (11) : 1346-1354. https://doi.org/10.1002/asi.24049.

Parsons, Mark A. et Peter A. Fox. 2013. « Is Data Publication the Right Metaphor ? ». *Data Science Journal* 12 : WDS32–WDS46. https://doi.org/10.2481/dsj.WDS-042.

Pärtel, Meelis. 2006. « Data Availability for Macroecology : How to Get More out of Regular Ecological Papers ». *Acta Oecologica* 30 (1) : 97-99. https://doi.org/10.1016/j.actao.2006.02.002.

Pasquetto, Irene V., Bernadette M. Randles et Christine L. Borgman. 2017. « On the Reuse of Scientific Data ». *Data Science Journal* 16. https://doi.org/10.5334/dsj-2017-008.

Passeron, Jean-Claude. 2011. *Le raisonnement sociologique : un espace non poppérien de l'argumentation*. Paris : Albin Michel.

Patil, Prasad, Roger D. Peng et Jeffrey T. Leek. 2016. « A Statistical Definition for Reproducibility and Replicability ». *BioRxiv*. https://doi.org/10.1101/066803.

Peels, Rick et Lex Bouter. 2018. « Humanities Need a Replication Drive too ». *Nature* 558 (7710): 372-373. https://doi.org/10.1038/d41586-018-05454-w.

Peng, Roger D., Francesca Dominici et Scott L. Zeger. 2006. « Reproducible Epidemiologic Research ». *American Journal of Epidemiology* 163 (9) : 783-789. https://doi.org/10.1093/aje/kwj093.

Peng, Roger D. et Stephanie C. Hicks. 2021. « Reproducible Research : A Retrospective ». *Annual Review of Public Health* 42 (1) : 79-93. https://doi.org/10.1146/annurev-publhealth-012420-105110.

Penev, Lyubomir, Vishwas Chavan, Teodor Georgiev et Pavel Stoev. 2012. « Data Papers as Incentives for Opening Biodiversity Data: One Year of Experience and Perspectives for the Future ». Poster présenté à *EU BON : Building the European Biodiversity Observation Network*. https://pensoft.net/img/upl/file/DataPaperPoster.pdf.

Peters, Isabella, Peter Kraker, Elisabeth Lex, Christian Gumpenberger et Juan Gorraiz. 2016. « Research Data Explored : An Extended Analysis of Citations and Altmetrics ». *Scientometrics* 107 (2) : 723-744. https://doi.org/10.1007/s11192-016-1887-4.

Pigenet, Yaroslav et Lydia Ben Ytzhak. 2014. « Fraude : mais que fait la recherche ? ». *CNRS Le journal* 278 : 17. https://lejournal.cnrs.fr/sites/default/files/numeros_papier/jdc_278_complet_bd3_0.pdf.

Pignard-Cheynel, Nathalie. 2004. « La communication des sciences sur Internet : stratégies et pratiques ». Thèse de doctorat, université Stendhal Grenoble 3. https://cds.cern.ch/record/813625.

Pinche, Ariane. 2021. *CREMMA Medieval, an Old French dataset for HTR and segmentation* (v1.0.1 Bicerin). https://doi.org/10.5281/zenodo.5235186.

Pinche, Ariane et Thibault Clérice. 2021. *HTR-United/cremma-medieval : 1.0.1 Bicerin (DOI)* (v1.0.1). Zenodo. https://doi.org/10.5281/ZENODO.5235186.

Pinon, Pierre. 2012. *Atlas du Paris haussmannien. La ville en héritage du Second Empire à nos jours*. Paris : Parigramme.

Piwowar, Heather A. et Todd J. Vision. 2013. « Data Reuse and the Open Data Citation Advantage». *PeerJ* 1 : e175. https://doi.org/10.7717/peerj.175.

Plesser, Hans E. 2018. « Reproducibility vs. Replicability : A Brief History of a Confused Terminology ». *Frontiers in Neuroinformatics* 11 : 76. https://doi.org/10.3389/fninf.2017.00076.

Pletschacher, Stefan et Apostolos Antonacopoulos. 2010. « The PAGE (Page Analysis and Ground-Truth Elements) Format Framework ». Communication présentée à *20th International Conference on Pattern Recognition*. https://doi.org/10.1109/ICPR.2010.72.

Pontille, David. 2007. « Matérialité des écrits scientifiques et travail de frontières : le cas du format IMRAD », dans *Sciences et frontières*, édité par Philippe Hert et Marcel Paul-Cavallier, 229-253. Louvain-la-Neuve : EME Éditions. https://halshs.archives-ouvertes.fr/halshs-00268991.

Pouyllau, Stéphane. 2011. « ISIDORE : une plateforme de recherche de documents et d'information pour les Sciences Humaines et Sociales ». *La lettre de l'inshs* 11. https://archivesic.ccsd.cnrs.fr/sic_00605642.

Pouyllau, Stéphane. 2014. « Huma-Num, les humanités numériques en réseau ». *Arabesques* 74 : 13-15. https://publications-prairial.fr/arabesques/index.php?id=944.

Pouyllau, Stéphane. 2016. « Isidore Suggestion, des recommandations de lecture pour les blogs de science ». *I2D - Information, données & documents* 53 (2) : 44. https://doi.org/10.3917/i2d.162.0044.

Pouyllau, Stéphane. 2021. « Isidore Suggestion, des recommandations de lecture pour les blogs de science ». I2D – Information, données & documents, 2016, Web de données et création de valeurs : le champ des possibles, 53 (2). https://archivesic.ccsd.cnrs.fr/sic_01348561v1.

Pouyllau, Stéphane, Laurent Capelli, Jean-Luc Minel, Mélanie Bunel, Nicolas Sauret, Olivier Baude, Hélène Jouguet, Pauline Busonera et Adrien Desseigne. 2021. « ISIDORE a 10 ans ». *Zenodo*. https://doi.org/10.5281/zenodo.5699997.

Pouyllau, Stéphane, Jean-Luc Minel, Shadia Kilouchi et Laurent Capelli. 2012. « Bilan 2011 de la plateforme ISIDORE et perspectives 2012-2015 ». Comité de pilotage du TGE Adonis. https://archivesic.ccsd.cnrs.fr/sic_00690558.

R Core Team. 2022. « R : A Language and Environment for Statistical Computing ». Vienne : R Foundation for Statistical Computing. https://www.R-project.org/.

Ramos, Juan. 2003. « Using TF-IDF to Determine Word Relevance in Document Queries ». Dans *Proceedings of the first instructional conference on machine learning*, 242 : 29-48. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c.

Rebouillat, Violaine. 2019. « Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs ». Thèse de doctorat, Conservatoire national des arts et métiers (CNAM). https://hal.archives-ouvertes.fr/tel-02447653.

Rees, Jonathan. 2010. Recommendations for Independent Scholarly Publication of Data Sets. San Francisco, CA : Creative Commons.

Renneville, Marc et Stéphane Pouyllau. 2015. « Huma-Num. La TGIR des humanités numériques. Rapport d'activité 2013-2015 ». Technical Report. TGIR Huma-Num

(UMS 3598). https://halshs.archives-ouvertes.fr/halshs-01390938.

Renouf, Antoinette. 1984. « Corpus Development at Birmingham Universit ». *Costerus New Series Online* 45 : 3-39.

Rentier, Bernard. 2018. *science ouverte, le défi de la transparence*. Bruxelles : Académie Royale de Belgique.

Reul, Christian, Christoph Wick, Maximilian Nöth, Andreas Büttner, Maximilian Wehner et Uwe Springmann. 2021. « Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning ». Communication présentée à *ACM Conference (HIP'21)*, 6. New York, NY, USA : ACM. http://arxiv.org/abs/2106.07881.

Revel, Jacques. 2012. « Les sciences historiques ». Dans *Épistémologie des sciences sociales* édité par Jean-Michel Berthelot, 19-76. Paris : PUF. https://doi.org/10.3917/puf.berth.2012.01.0019.

Rey-Coyrehourcq, Sébastien, Robin Cura, Laure Nuninger, Julie Gravier, Lucie Nahassia et Ryma Hachi. 2017. Vers une recherche reproductible dans un cadre interdisciplinaire : enjeux et propositions pour le transfert du cadre conceptuel et la réplication des modèles. Tours : Presses universitaires François Rabelais. https://hal.archives-ouvertes.fr/hal-01677950.

Reymonet, Nathalie. 2017. « Améliorer l'exposition des données de la recherche : la publication de *data papers* ». https://archivesic.ccsd.cnrs.fr/sic_01427978.

Rhys, Hefin I. 2020. Machine Learning with R, the tidyverse, and mlr. Shelter Island : Manning Publications.

Ribary, Marton. 2020a. « A Relational Database of Roman Law Based on Justinian's Digest ». *Journal of Open Humanities Data* 6 (1) : 5. https://doi.org/10.5334/johd.17.

Ribary, Marton. 2020b. « A Relational Database of Roman Law based on Justinian's Digest ». *Figshare*. https://doi.org/10.6084/m9.figshare.12333290.v2.

Ribary, Marton. 2020c. *pyDigest: A GitLab repository of scripts, files and documentation*. GitLab.

Ribary, Marton. 2021. « Open Research Methods for a Digital Turn in Roman Law » présenté à Surrey Open Research and Transparency Showcase 2021. https://www.youtube.com/watch?v=FiVU8FbD6YI&list=PLlVZ6vn-g5xb_il3e-bxptjBo8nHdOVD2X&index=11.

Ribary, Marton et Barbara McGillivray. 2020. « A corpus approach to Roman Law based on Justinian's Digest ». *Informatics* 7 (44). https://doi.org/10.3390/informatics7040044.

Richou, Louki-Géronimo et Joachim Schöpfel. 2023. « L'option d'un entrepôt national ». Dans *Partage et valorisation des données de la recherche. Développements, tendances et modèles*, édité par Joachim Schöpfel et Violaine Rebouillat, 127-146. Londres, ISTE.

Rijcke, Sarah de et Bart Penders. 2018. « Resist Calls for Replicability in the Humanities ». *Nature* 560 (7716) : 29. https://doi.org/10.1038/d41586-018-05845-z.

Rivet, Alain, Marie-Laure Bachèlerie, Auriane Denis-Meyere et Delphine Tisserand. 2018. *Traçabilité des activités de recherché et gestion des connaissances – Guide pratique de mise en place*. Paris : Mission pour les initiatives transverses et interdisciplinaires du CNRS. https://qualite-en-re-

cherche.cnrs.fr/wp-content/uploads/2021/08/guide_tracabilite_activites_recherche_gestion_connaissances.pdf.

Robinson-García, Nicolas, Evaristo Jiménez-Contreras et Daniel Torres-Salinas. 2016. « Analyzing Data Citation Practices using the Data Citation Index ». *Journal of the Association for Information Science and Technology* 67 (12) : 2964–2975. https://doi.org/10.1002/asi.23529.

Rowhani-Farid, Anisa, Michelle Allen and Adrian G. Barnett. 2017. « What Incentives Increase Data Sharing in Health and Medical Research ? A Systematic Review ». *Research Integrity and Peer Review* 2 (1) : 4. https://doi.org/10.1186/s41073-017-0028-9.

RStudio Team. 2022. « RStudio : Integrated Development for R ». Boston : RStudio, Inc. http://www.rstudio.com/.

Ruggles, Steven. 2021. « The Revival of Quantification : Reflections on Old New Histories ». *Social Science History* 45 (1) : 1-25. https://doi.org/10.1017/ssh.2020.44.

Schieder, Wolfgang. 1963. *Anfänge der deutschen Arbeiterbewegung. Die Auslandsvereine im Jahrzehnt nach der Julirevolution von 1830. Industrielle Welt* 4. Stuttgart : Klett.

Schlagdenhauffen, Régis. 2020. « Optical Recognition Assisted Transcription with Transkribus : The Experiment Concerning Eugène Wilhelm's Personal Diary (1885-1951) ». *Journal of Data Mining and Digital Humanities* Atelier Digit_Hum. https://doi.org/10.46298/jdmdh.6249.

Schmidt, Frank L. et In-Sue Oh. 2016. « The Crisis of Confidence in Research Findings in Psychology : Is Lack of Replication the Real Problem ? Or Is It Something Else ? ». *Archives of Scientific Psychology* 4 (1) : 32-37. https://doi.org/10.1037/arc0000029.

Schöpfel, Joachim. 2020. « À propos des données de recherche en SHS ». *La Lettre de l'InSHS* 63 : 24–26. https://www.inshs.cnrs.fr/sites/institut_inshs/files/download-file/lettre_infoINSHS_63v3_0.pdf.

Schöpfel, Joachim, Dominic Farace, Hélène Prost et Antonella Zane. 2019. « Data Papers as a New Form of Knowledge Organization in the Field of Research Data». *Knowledge Organization* 46 (8) : 622–638. https://doi.org/10.5771/0943-7444-2019-8-622.

Schöpfel, Joachim, Farace, Dominic J., Prost, Hélène et Antonella Zane. 2019. « Data Papers as a New Form of Knowledge Organization in the Field of Research Data ». Dans *12e colloque international d'ISKO-France. Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?* Montpellier. https://shs.hal.science/halshs-02284548.

Schöpfel, Joachim, Dominic Farace, Hélène Prost, Antonella Zane et Birger Hjorland. 2020. « Data Documents». dans *Encyclopedia of Knowledge Organization* édité par Birger Hjørland et Claudio Gnoli. https://www.isko.org/cyclo/data_documents.

Schöpfel, Joachim, Éric Kergosien, Stéphane Chaudiron, Bernard Jacquemin et Hélène Prost. 2022. « Les revues SIC face à l'enjeu de la transparence et de l'ouverture. Une étude empirique ». *8e Conférence Document Numérique et Société. Liège, Belgique.* https://hal.science/hal-03748627v1 Yoon, Ayoung. 2014. « End Users' Trust in Data Repositories : Definition and Influences on Trust Development. *Archival Science*, 14 (1) : 17-34. https://doi.org/10.1007/s10502-013-9207-8.

Schöpfel, Joachim et Violaine Rebouillat (dir.). 2023. *Partage et valorisation des données de la recherche*. London : ISTE Editions.

Schreibman, Susan, Ray Siemens et John Unsworth, éd. 2004. *A Companion to Digital Humanities*. Oxford : Blackwell Publishing.

Sebastiani, Fabrizio. 2002. « Machine Learning in Automated Text Categorization ». ACM Computing Surveys 34 (1) : 1-47. https://doi.org/10.1145/505282.505283.

Senftleben, Martin. 2022. « Study on EU Copyright and Related Rights and Access to and Reuse of Data ». *Publications Office of the European Union*. http://doi.org.org/10.2777/78973.

Seo, Sunkyung et Jihyun Kim. 2020. « Data Journals : Types of Peer Review, Review Criteria, and Editorial Committee Members' Positions ». *Science Editing* 7 (2) : 130–135. https://doi.org/10.6087/kcse.207.

Shotton, David. 2009. « *Semantic Publishing : The Coming Revolution in Scientific Journal Publishing* ». *Learned Publishing* 22 (2) : 85–94. https://doi.org/10.1087/2009202

Smith, Mackenzie. 2011. « Data Papers in the Network Era». *Something's Gotta Give* édité par Beth R. Bernhardt, Leah H. Hinds et Katina P. Strauch. West Lafayette : Purdue University Press. https://doi.org/10.5703/1288284314871.

Solopova, Veronika, Tatjana Scheffler et Mihaela Popa-Wyatt. 2021. « A Telegram Corpus for Hate Speech, Offensive Language, and Online Harm ». *Journal of Open Humanities Data* 7 : 9. https://doi.org/10.5334/johd.32.

Souza, Allan Rocha de. 2020 « Covid-19, Text and Data Mining and Copyright : The Brazilian Case ». Dans *WIPO-WTO Colloquium Papers Volume 11*, édité par Yogesh Pai, Jessyca van Weelde and Wardi Zaman, 1-14. WIPO Academy and WTO IP, Government Procurement and Competition Division.

Souza, Allan Rocha de, Luca Schirru et Miguel Bastos Alvarenga. 2020. « Copyright and Data ans Text Mining in the Fight against Covid-19 in Brazil ». *Liinc em Revista* 16(2) : e5536.

Staff, Science. 2011. « Challenges and Opportunities ». *Science* 331 (6018) : 692-693. https://doi.org/10.1126/science.331.6018.692.

Star, Susan Leigh. 2010. « Ceci n'est pas un objet-frontière ! ». *Revue d'anthropologie des connaissances* 4 (1). https://doi.org/10.3917/rac.009.0018.

Stiglitz, Joseph E. 1999. « Knowledge as a Global Public Good ». Dans *Global Public Goods : International cooperation in the 21st Century*, édité par Inge Kaul, Isabelle Grunberg, Marc Stern, 308-325. Oxford : Oxford University Press. https://doi.org/10.1093/0195130529.003.0015.

Stodden, Victoria, Friedrich Leisch et Roger D Peng, éd. 2014. *Implementing Reproducible Research*. Boca Raton : CRC Press.

Stokes, Peter A., Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot et El Hassane Gargem. 2021. « The eScriptorium VRE for Manuscript Cultures ». *Classics@ Journal* 18 (1). https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/.

Stökl Ben Ezra, Daniel. 2021. « L'infrastructure eScriptorium de reconnaissance automatique d'écriture manuscrite (HTR) ». Communication présentée à *Rendez-vous IIIF360 2021*. https://projet.biblissima.fr/fr/infrastructure-es-

criptorium-reconnaissance-automatique-ecriture-manuscrite-htr.

Straka, Milan et Jana Straková. 2017. « Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe ». Dans *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, 88-99. Vancouver : Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3009.

Stutzmann, Dominique. 2011. « Paléographie statistique pour décrire, identifier, dater… Normaliser pour coopérer et aller plus loin ? » In *Kodikologie und Paläographie im digitalen Zeitalter = Codicology and Palaeography in the Digital Age*, édité par Franz Fischer, Christiane Fritze et Georg Vogeler, BoD, 2 : 247-277. Schriften des Instituts für Dokumentologie und Editorik. BoD. https://halshs.archives-ouvertes.fr/halshs-00596970.

Stutzmann, Dominique, Jean-François Moufflet et Sébastien Hamel. 2017. « La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique ». *Médiévales. Langues, Textes, Histoire* 73 : 67-96. https://doi.org/10.4000/medievales.8198.

Szabo, Dimitri. 2019. « Un aperçu de Dataverse ». *2e Réunion annuelle 2019 du nœud national RDA France* Research Data Alliance France, Paris. http://dx.doi.org/10.17180/xwz6-bt15.

Tedersoo, Leho, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, *et al.* 2021. « Data Sharing Practices and Data Availability upon Request Differ Across Scientific Disciplines ». *Scientific Data* 8 (1) : 192. https://doi.org/10.1038/s41597-021-00981-0.

Teklia. 2021. *Teklia/Arkindex : 0.15.4* (0.15.4). https://teklia.com/solutions/arkindex/releases/0-15-4/.

Thelwall, Mike. 2020. « Data in Brief : Can a Mega-Journal for Data be Useful ? » *Scientometrics* 124 (1) : 697–709. https://doi.org/10.1007/s11192-020-03437-1

Thillay, Alain. 1999. « Les artisans étrangers au faubourg Saint-Antoine à Paris (1650–1793) ». In *Les Étrangers dans la ville. Minorités et espaces urbains du Moyen Âge à l'époque moderne*, édité par Jacques Bottin et Donatella Calabi, 261-269. Paris : Éditions de la Maison des sciences de l'Homme.

Van Spanje, Joost. 2018. *Controlling the Electoral Marketplace. How Established Parties Ward Off Competition*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-58202-3.

Vines, Timothe H., Arianne Y.K. Albert, Rose L. Andrew, Florence Débarre, *et al.* 2014. « The Availability of Research Data Declines Rapidly with article Age ». *Current Biology* 24 (1) : 94-97. https://doi.org/10.1016/j.cub.2013.11.014.

Wacquet, Françoise. 2022. Dans les coulisses de la science. Petites mains et autres travailleurs invisibles. Paris : CNRS Éditions.

Walters, William H. 2020. « Data Journals : Incentivizing Data Access and Documentation within the Scholarly Communication System ». *Insights the UKSG Journal* 33 : 1–20. https://doi.org/10.1629/uksg.510.

Wang, Xiaoguang, Qingyu Duan et Mengli Liang. 2021. « Understanding the Process of Data Reuse : An Extensive Review». *Journal of the Association for Information Science and Technology* 72 (9) : 1161–1182. https://doi.org/10.1002/asi.24483.

Welbers, Kasper, Wouter Van Atteveldt et Kenneth Benoit. 2017. « Text Analysis in R ». *Communication Methods and Measures* 11 (4) : 245-265. https://doi.org/10.1080/19312458.2017.1387238.

Werner, Michael. 1995. « Étrangers et immigrants à Paris autour de 1848 : l'exemple des Allemands ». Dans *Paris und Berlin in der Revolution 1848*, édité par Ilja Mieck, Horst Möller, et Jürgen Voss, 199-213. Sigmaringen : Thorbecke.

Wickham, Hadley. 2011. « ggplot2 ». *Wiley Interdisciplinary Reviews : Computational Statistics* 3 (2) : 180-185. https://doi.org/10.1002/wics.147.

Wieling, Martijn, Josine Rawee et Gertjan van Noord. 2018. « Reproducibility in Computational Linguistics : Are We Willing to Share ? » *Computational Linguistics* 44 (4): 641-649. https://doi.org/10.1162/coli_a_00330.

Wigdorowitz, Mandy, Barbara McGillivray, Andrea Farina, Simon Hengchen, Nilo Pedrazzini, Iacopo Ghinassi et Marton Ribary. 2022. *JOHD Data Paper Template*. https://www.overleaf.com/latex/templates/johd-data-paper-template/mqcypcbntsds.

Wijffels, Jan, Milan Straka et Jana Straková. 2023. « Udpipe : Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the "UDPipe" "NLP" Toolkit ». https://cran.r-project.org/web/packages/udpipe/index.html.

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, *et al*. 2016. « The FAIR Guiding Principles for Scientific Data Management and Stewardship ». *Scientific Data* 3 (1) : 160018. https://doi.org/10.1038/sdata.2016.18.

World Trade Organization, Agreement on Trade-Related Aspects of Intellectual Property Rights (15 April 1994, Annex 1C) 1869 UNTS. 299 (TRIPS Agreement),

Xia, Yufei, Chuanzhe Liu, YuYing Li et Nana Liu. 2017. « A Boosted Decision Tree Approach Using Bayesian Hyper-Parameter Optimization for Credit Scoring ». *Expert Systems with Applications* 78 : 225-241. https://doi.org/10.1016/j.eswa.2017.02.017.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr.* 2e éd. Boca Raton : CRC Press.

Yoon, JungWon, EunKyung Chung, Jae Yun Lee et Jihyun Kim. 2019. « How Research Data is Cited in Scholarly Literature: A Case Study of HINTS ». *Learned Publishing* 32 (3) : 199-206. https://doi.org/10.1002/leap.1213.

Zech, Herbert. 2016. « A Legal Framework for a Data Economy in the European Digital Single Market : Rights to Use Data ». *Journal of Intellectual Property Law & Practice* 11(6) : 460-470. https://doi.org/10.1093/jiplp/jpw049.

Zimmerman, Ann. 2007. « Not by Metadata Alone : the Use of Diverse Forms of Knowledge to Locate Data for Reuse ». *International Journal on Digital Libraries* 7 (1-2) : 5–16. https://doi.org/10.1007/s00799-007-0015-8.

Biographies

1. Professeur de phonétique à l'UQAC depuis 2008, **Vincent Arnaud** est détenteur d'un doctorat en cotutelle entre l'Université de Franche-Comté (France) et l'Université Laval (Canada), il collabore à différents projets en bioacoustique, en traitement automatique des langues et en apprentissage machine.
2. Professeur en informatique à l'UQAC depuis 2016, **Kévin Bouchard** est détenteur d'un doctorat de l'Université du Québec en Outaouais et a effectué un postdoctorat à l'université de Californie à Los Angeles. Il travaille sur les habitats intelligents et les technologies pour la santé en appliquant son expertise en intelligence artificielle.
3. Issue du monde de la recherche en biologie et écologie marine, **Mélanie Bunel** a notamment travaillé pendant près de trois ans en Nouvelle-Calédonie pour l'Université de Nouméa puis pour le Secrétariat du Pacifique Sud pour des programmes de recherche sur des thématiques de gestion et protection des milieux marins et d'économie de la pêche. Elle a intégré Huma-Num en 2020 en tant qu'ingénieure d'études spécialisée dans les domaines de l'ingénierie documentaire pour le projet européen TRIPLE. Elle rejoint le HN Lab en 2021 pour travailler sur le projet Huma-Num science ouverte (HNSO).
4. **Alix Chagué** est doctorante en Humanités Numériques à l'Université de Montréal et à l'École Pratique des Hautes Études, membre de l'équipe ALManACH d'Inria Paris. Ses recherches portent principalement sur la transcription automatique des documents manuscrits et son intégration dans des chaînes de traitement pour l'édition numérique ou la valorisation de collections.
5. **Thibault Clérice** est docteur en Lettres et Civilisation Antiques (Laboratoire HISOMA, Université Lyon 3). Ses recherches portent principalement sur le traitement automatique des langues anciennes à travers l'apprentissage profond, la mise à disposition de corpora, la recherche en méthodes computationnelles appliquées aux humanités.
6. **Clémentine Cottineau-Mugadza** est géographe, chargée de recherche CNRS détachée et assistant professor of Urban Studies à la Technische Universiteit de Delft aux Pays-Bas. Ses travaux portent sur l'analyse et la modélisation des inégalités socio-économiques dans et entre les villes.
7. **Alina Danciu** est responsable de l'équipe « Documentation-Diffusion » du CDSP.
8. **Anna Egea** est ingénieure de recherche, documentaliste au Centre de sociologie des organisations.
9. **Guillaume Garcia** est ingénieur de recherche, pôle « Méthodes et valorisation » du CDSP.

10. **Victor Gay** est enseignant-chercheur. Ses travaux sont à la croisée de l'histoire économique, de l'économie du travail et de l'économie de la culture. Il a notamment travaillé sur les conséquences de la Première Guerre mondiale sur le travail des femmes.

11. **Cyril Heude** est data librarian à Sciences Po.

12. **Gérald Kembellec** est maître de conférences en sciences de l'information et de la communication au Cnam, il anime la thématique « Data, médiation, valorisation » du laboratoire « Dispositifs d'information et de communication à l'ère du Numérique ». Il était détaché au département des humanités numériques de l'IHA en 2020 et 2021.

13. Titulaire d'un doctorat en sciences de l'information de l'université de Michigan à Ann Arbor, **Jihyun Kim** est professeure en sciences de l'information et de la bibliothèque au Department of Library and Information Science, Ewha Womans University, Séoul.

14. **Mareike König** est directrice adjointe et directrice du département Humanités numériques à l'Institut historique allemand (IHA). Elle est à l'origine du projet « Adressbuch der Deutschen in Paris von 1854 ». Historienne de l'histoire franco-allemande du XIX^e^ siècle, ses recherches portent également sur l'Histoire à l'ère numérique et sur la communication scientifique dans les médias sociaux.

15. **Christine Kosmopoulos** est ingénieure de recherche au CNRS et responsable éditoriale de la revue européenne de géographie *Cybergeo*. Ses travaux de recherche portent sur l'impact des nouvelles technologies de l'information et de la communication en SHS notamment dans le cadre de la science ouverte.

16. **Victoria Le Fourner** est spécialiste du traitement des données en SHS. Elle a exercé en tant qu'ingénieure d'étude au CNRS et à la BULAC. Depuis 2024, elle est consultante en gouvernance des données au sein du cabinet TNP.

17. **Alicia León y Barella** est conservatrice des bibliothèques et cheffe du service science ouverte du SCD de l'Université de Lille.

18. **Thomas Margoni** is Research Professor of Intellectual Property Law at the Faculty of Law and Criminology, KU Leuven, and member of the Board of Directors at the Centre for IT & IP Law (CiTiP). At CiTiP, he is PI for several projects in the field of IP, Data and open science, including Skills4EOSC, Data Spaces Support Centre, Recreating.eu, XRECO and DAFNE+.

19. **Paola Marongiu** est doctorante en linguistique à l'Université de Neuchâtel et rédactrice de réseaux sociaux du *Journal of Open Humanities Data* ; ses recherches portent sur la sémantique, notamment l'expression de la modalité, le latin et les humanités numériques.

20. **Barbara McGillivray** est maître de conférences en humanités numériques et *cultural computing* au King's College de Londres, Turing fellow et rédactrice en chef du *Journal of Open Humanities Data* ; ses recherches

portent sur la linguistique computationnelle des textes historiques.

21. Étudiant au baccalauréat en informatique à l'UQAC depuis 2020, **Gilles-Philippe Morin** est détenteur d'un baccalauréat en sciences de la santé de l'Université de Sherbrooke ; il s'intéresse aux neurosciences, à la linguistique et plus particulièrement au traitement automatique des langues.

22. Ingénieure en poste au GBIF France (point nodal français du réseau international GBIF) depuis 2013, **Sophie Pamerlon** est chargée de la standardisation, de la diffusion et de la mise en valeur des données de biodiversité dans le contexte de la science ouverte et des principes FAIR, notamment via la promotion des standards d'échange de données et de métadonnées, des Data Papers et des bonnes pratiques liées au dépôt, à la description et au partage des données pour en améliorer l'accessibilité et la réutilisation.

23. **Christophe Parisse** est linguiste, spécialiste du développement du langage chez l'enfant et de la linguistique de corpus.

24. **Nilo Pedrazzini** est assistant de recherche à l'Alan Turing Institute, Londres ; il est aussi assistant de rédaction pour le *Journal of Open Humanities Data*. Ses recherches portent sur les humanités numériques et la sémantique formelle pour les langues historiques.

25. Ingénieur de recherche au CNRS, **Stéphane Pouyllau** est spécialisé depuis 1999 dans les humanités numériques et l'informatisation des données en sciences humaines et sociales. Cocréateur en 2005 du Centre national pour la numérisation de sources visuelles, il est aussi l'initiateur de MédiHAL, archive ouverte pour photographies et images scientifiques. De 2009 à 2011, il a codirigé la plateforme Isidore au TGE Adonis, avant de devenir directeur de Corpus-IR et cofondateur en 2013 de la TGIR Huma-Num, où il est actuellement directeur adjoint. Lauréat du Cristal du CNRS en 2009, ses travaux sont accessibles sur ISIDORE.

26. **Denise Pumain** est professeure émérite de géographie à l'Université Paris I Panthéon-Sorbonne. Spécialiste des théories et modèles de systèmes urbains, elle a reçu le prix Vautrin Lud en 2010. Elle a créé en 1996 la première revue en ligne de géographie, *Cybergeo, revue européenne de géographie*.

27. Docteure en Sciences de l'information et de la communication, **Violaine Rebouillat** consacre ses recherches à la communication scientifique, étudiant en particulier les pratiques de production, d'utilisation et de partage des données de recherche.

28. **Marton Ribary** est maître de conférences au département de droit et criminologie de la Royal Holloway (Université de Londres). Ses recherches portent sur le raisonnement juridique et sur sa modélisation informatique dans le droit privé, tant ancien (romain et rabbinique) que moderne (anglais). Il est éditeur associé et assistant de rédaction pour le *Journal of Open Humanities Data*.

29. **Alain Rivet** est responsable Qualité – Système d'information, Comité de pilotage du réseau Qualité en Recherche, membre des groupes de travail « Atelier-Données » et « Cahiers de laboratoire électroniques » du CNRS.

30. **Laurent Romary** était directeur de recherche à Inria jusqu'à sa nomination en tant que directeur à la culture et à l'information scientifique, en janvier 2022. Il a mené pendant des années des recherches en informatique linguistique et en humanités numériques, au cours desquelles il a développé un intérêt marqué pour la normalisation et le partage de données linguistiques ouvertes, et identifié la nécessité de mettre en œuvre des infrastructures mutualisées au service de la science ouverte.

31. **Nicolas Sauret** est maître de conférences en Sciences de l'information et de la communication à l'Université Paris 8 Vincennes - Saint-Denis. Ses recherches portent sur la nature collective de la production, de la circulation et de la légitimation des connaissances. Entre recherche théorique et expérimentations éditoriales, ses travaux s'ancrent dans une recherche-action qu'il pratique avec ses étudiantes et étudiants au sein de cours ou dans divers contextes participatifs, explorant ainsi les pratiques d'écritures numériques et les nouvelles fabriques du savoir, qu'elles soient ouvertes, collectives ou participatives. | nicolassauret.net

32. **Luca Schirru** is Postdoctoral researcher at Centre for IT & IP Law (CiTiP), Faculty of Law and Criminology, University of Leuven (KU Leuven), Belgium ; luca.schirru@kuleuven.be. His research is funded under the Skills4EOSC project. Skills4EOSC has received funding from the European Union's Horizon Europe research and innovation Programme under Grant Agreement No. 101058527.

33. **Joachim Schöpfel** est maître de conférences en sciences de l'information et de la communication à l'Université de Lille et membre du laboratoire GERiiCO. Son domaine d'expertise est l'information et la communication scientifique. Il est consultant indépendant du cabinet Ourouk.

34. **Brad Spitz** et Docteur en droit et avocat au Barreau de Paris – Realex.

35. **Evan Virevialle** est étudiant en 2e année du Master communication du savoir, technologies de la connaissance et management de l'information à l'université Paris 1 Panthéon-Sorbonne. Il est également assistant de recherche à l'IHA, il y collabore notamment au projet « Adressbuch der Deutschen in Paris von 1854 » sur les aspects de qualification et d'enrichissement des données, d'interface et de cartographie.

Résumés

Le paysage des *data papers*

Victoria Le Fourner et Joachim Schöpfel

Ce premier chapitre dresse un panorama général des articles de données (*data papers*) en s'appuyant sur les observations et les pratiques dans les différentes disciplines et non exclusivement en SHS. Dans un premier temps, il propose une définition du concept et décrit le développement des *data papers* ainsi que des revues de données (*data journals*). Dans un second temps, il analyse la fonction, le contenu et la structure de ces articles, avec un point sur l'accès aux données et sur leur réutilisation. Enfin, la troisième partie du chapitre aborde la production des *data papers*, le processus, en mettant l'accent sur la question de la qualité (sélection, évaluation) et de l'impact.

Révéler les formes et logiques de citation des *data papers* en archéologie : le cas du *Journal of Open Archaeology Data*

Violaine Rebouillat

Le présent chapitre investit la question des pratiques de citation des *data papers* en archéologie à travers le cas d'étude du *Journal of Open Archaeology Data* (*JOAD*). En ciblant les auteurs ayant cité un des *data papers* de la revue, l'approche choisie d'une enquête par questionnaire permet de révéler leurs motivations. À partir des réponses de 19 auteurs, l'enquête montre que la citation des *data papers* en archéologie est le fruit de chercheurs sensibles à l'ouverture des résultats de la science et s'inscrit dans un contexte d'émergence de ce nouveau type de publication. Souvent cité à la place du jeu de données, le *data paper* est perçu à la fois comme un point d'entrée vers celui-ci et comme un moyen de délester l'article de recherche d'une partie méthodologique sur le processus de collecte des données. Ces résultats révèlent donc une utilisation du *data paper* dans les pratiques de citation comme document situé à l'articulation entre données et article de recherche.

Le *Journal of Open Humanities Data* (*JOHD*) : enjeux et défis dans la publication de *data papers* pour les sciences humaines et sociales (SHS)

Paola Marongiu, Nilo Pedrazzini, Marton Ribary et Barbara McGillivray

Dans ce chapitre, nous présentons le *data paper*en tant que forme de publication qui vise à valoriser le travail de préparation et le traitement des données, en adoptant la perspective des sciences humaines et sociales (SHS). Après avoir reconstruit une définition aussi complète que possible de *data paper*, nous analyserons son évolution au fil du temps. Nous essaierons d'identifier les obstacles qui ont pu influencer la moindre fortune des *data paper*sen SHS par rapport aux sciences dures jusqu'à présent, en attirant l'attention sur l'extrême hétérogénéité des sujets de recherche et du type de données en SHS. Ensuite, nous présentons le *Journal of Open Humanities Data*(*JOHD*), qui se consacre à la publication d'articles axés sur les données pour les SHS. Nous partagerons notre expérience dans ce milieu, en présentant également une enquête sur l'impact des *data papers* sur la réutilisation des données qu'ils décrivent. Pour conclure, nous proposerons un modèle pyramidal, dans lequel le *data paper* participe à la valorisation des résultats de recherche, ainsi que du

travail de curation et analyse des données, au vu des valeurs de science ouverte et de partage des données.

Un *data paper* en SHS : pourquoi, pour qui, comment ?

Victor Gay

Destinée aux chercheurs en sciences humaines et sociales souhaitant se lancer dans l'écriture d'un *data paper*, ce chapitre propose un retour d'expérience prenant appui sur la production récente d'un *data paper* (Gay, 2021) et aborde les enjeux auxquels un auteur de *data paper* est souvent confronté. Tout d'abord, pourquoi écrire un *data paper* ? Alors que la valeur scientifique de la production de données reste peu reconnue, ce format éditorial constitue un outil qui peut permettre aux producteurs de données de faire reconnaître leur contribution scientifique. Ensuite, pour qui écrire un *data paper* ? Alors que seules quelques revues en sciences humaines et sociales acceptent aujourd'hui ce genre d'article, cette rareté peut paradoxalement constituer une chance pour les producteurs de données dans la mesure où cela peut leur permettre d'atteindre un lectorat relativement large et interdisciplinaire. Enfin, comment écrire un *data paper* afin qu'il constitue une véritable clé d'accès pour la compréhension et la réutilisation des données décrites ? Il semble opportun ici de s'inspirer de modèles éprouvés issus des sciences dures, tout en les adaptant aux spécificités des sciences humaines et sociales.

Data paper en humanités numériques : Adressbuch 1854

Mareike König, Gérald Kembellec et Evan Virevialle

Ce chapitre présente un *data paper* qui porte sur les jeux de données liés au projet « *Adressbuch der Deutschen in Paris von 1854* » au long cours d'histoire sur l'immigration des Allemands à Paris au XIX^{e} siècle réalisé à l'Institut historique allemand de Paris. Le projet a pour objectif de mettre à la disposition des chercheuses et chercheurs et généalogistes, dans une interface, les informations sur les individus et entreprises allemands installés à Paris en 1854. Ces informations sont issues d'un document historique, une sorte de bottin de commerce ou pages jaunes des Allemands à Paris.
Outre la description de méthodes de collecte des données, ce chapitre présente la manière dont les jeux de données sont enrichis et valorisés dans le dispositif de consultation. Nous portons une attention particulière sur le respect des valeurs des humanités numériques et de la science ouverte, en cohérence avec les spécificités des historiennes et historiens, ainsi que les généalogistes amenés à le consulter.

Utiliser un *data paper* en traitement automatique des langues : un exemple de classification automatique de mémoires et de thèses universitaires

Vincent Arnaud, Kevin Bouchard et Gilles-Philippe Morin

Les notions de reproductibilité et de réplicabilité sont au centre de débats récents dans de nombreux champs disciplinaires. Les travaux en linguistique n'y font pas exception. Dans ce contexte, ce chapitre a pour objectif de partager une analyse menée en traitement automatique des langues avec

la création d'un article exécutable. Ce format de publication vise à partager chaque étape de l'analyse pour parvenir aux résultats obtenus, depuis les données jusqu'aux graphiques finaux. Cet article exécutable est disponible sous la forme d'un *notebook* dans lequel se combinent des informations méthodologiques et techniques, un accès aux données : figures, résultats, mais aussi le code informatique utilisé pour créer les graphiques et réaliser les analyses. Après avoir discuté de différents aspects de la notion de reproductibilité, seront indiqués le contexte, les choix méthodologiques et les résultats d'une expérience de classification automatique de textes universitaires dans différentes thématiques de recherche en fonction du vocabulaire utilisé par les auteurs de ces textes. La publication du *notebook* associé permet d'accéder au code informatique et aux détails méthodologiques et vise à favoriser la reproduction ou la réplication de cette expérience.

Une analyse des modèles et instructions des *data papers* : types d'informations contextuelles décrites par les *data journals*

Jihyun Kim

L'étude examine dans quelle mesure les composantes des *data papers* spécifiées par les revues représentent les types d'informations contextuelles nécessaires à la réutilisation des données. Une analyse du contenu de 15 modèles/instructions aux auteurs de *data papers* issue de 24 revues de données indexées par le *Web of Science* a été réalisée. Un schéma de codage a été élaboré sur la base d'études antérieures, comprenant quatre catégories : propriétés générales des données, informations sur la production des données, informations sur l'entrepôt et informations sur la réutilisation. Seuls quelques types d'informations contextuelles sont couramment demandées par les revues. Ces résultats suggèrent que les revues de données devraient fournir un ensemble plus standardisé des composantes du *data paper* afin d'apporter une information cohérente pour les réutilisateurs sur les données contextuelles pertinentes.

Évaluer un *data paper*, l'exemple de *Cybergeo*

Clémentine Cottineau-Mugadza, Christine Kosmopoulos et Denise Pumain

Ce chapitre relate la pratique d'évaluation des *data papers* d'après l'expérience de *Cybergeo*, une revue en ligne de géographie qui a créé sa rubrique en 2017. Le processus d'évaluation est adapté à ce type d'article pour valoriser l'étape de construction des données et pour en préserver la mémoire, tout en permettant la réutilisation des données. Cette ouverture contribue à la reproductibilité des travaux et à la cumulativité des connaissances. De nouveaux critères d'évaluation ont été ajoutés pour normaliser la présentation des articles afin de faciliter la compréhension de la construction des données et leur prise en main par de nouveaux utilisateurs. Certains ont trait à la forme des articles et sont communs aux *data papers* d'autres disciplines (description de la construction des données, archivage pérenne et reproductibilité). D'autres sont spécifiques à la nature et aux usages des données en géographie (comme l'échelle d'analyse, la granularité spatiale, etc.). Enfin, c'est aussi l'organisation de la revue qui a été adaptée pour garantir une évaluation rigoureuse et équitable de ce type d'articles.

La FAIRisation des données

Alain Rivet

La gestion rigoureuse et cohérente des données de la recherche constitue aujourd'hui un enjeu majeur pour la production de nouvelles connaissances scientifiques. Guidés par le « Plan National pour la science ouverte » qui prône la diffusion sans entrave des publications et des données de la recherche, les différents organismes de recherche s'emparent aujourd'hui de ces questions primordiales.

La science ouverte est une nouvelle approche de la démarche scientifique, basée sur la production collaborative des produits de la science, de leur partage, de leur libre circulation en vue de leur réutilisation. La Commission européenne a émis dès 2012 des recommandations concernant la diffusion des résultats scientifiques qui invitent les chercheurs à s'appuyer sur les principes *FAIR* (*Findable, Accessible, Interoperable, Reusable*) permettant de favoriser la découverte, l'accès, l'interopérabilité et la réutilisation des données.

La FAIRisation des données est un processus complexe, souvent long et coûteux qui nécessite des moyens techniques et humains. Gérer les données de la recherche comprend plusieurs étapes avant d'aboutir à la publication et l'archivage de données fiables, de qualité, respectueuses du droit des personnes et de la législation en vigueur. Nous essayerons d'apporter, dans ce chapitre, les informations permettant de répondre à ces objectifs.

Le rôle des licences dans la FAIRisation des données

Thomas Margoni, Luca Schirru et Brad Spitz

La FAIRisation des données et les licences « open source » telles que Creative Commons ont un dénominateur commun appliqué pour les données qui accompagnent le *data paper* : le droit d'auteur. En effet, la plupart des licences utilisées pour partager du contenu fonctionnent sur la base d'un droit d'auteur sous-jacent, et lorsque ce droit est absent (ou qu'une utilisation particulière se trouve en dehors du champ de l'exclusivité du droit d'auteur, comme dans le cas des exceptions et limitations), la licence n'est généralement pas activée. Ce point est central dans le cadre du débat sur la FAIRisation des données, dans la mesure où la plupart des données, comme les faits et les idées, ont été intentionnellement laissées en dehors du champ d'application du droit d'auteur. Dans ce chapitre, nous tenterons de présenter cette relation du point de vue du droit de l'UE, de façon concise, mais avec une base juridique solide, et nous envisagerons les domaines pouvant faire l'objet d'une réforme.

Science ouverte, plans de gestion de données et *data papers* au cœur d'une offre de services : l'exemple du SCD de l'université de Lille

Alicia León y Barella

Quelle est la place des *data papers* dans l'offre de services d'une bibliothèque universitaire ? Ce chapitre montre, à l'exemple du SCD de l'université de Lille, comment la formation aux articles de données s'articule avec les autres domaines d'appui à la recherche, notamment dans le cadre de l'atelier de la donnée, *Lille Open Research Data* (LORD) de l'écosystème Recherche Data Gouv.

Des corpus de langage oral aux *data papers*

Christophe Parisse

Les corpus de langage oral sont des données rares, coûteuses et précieuses pour la recherche en sciences du langage. Ces

caractéristiques ont depuis longtemps incité les chercheurs à les partager et à les proposer à toute la communauté scientifique. Ils représentent des données qui ont, depuis longtemps, été décrites dans des articles scientifiques, et qui sont distribuées de manière accessible pour d'autres recherches et travaux. Le couple « données et description » ressemble fortement au couple formé aujourd'hui par les données et les *data paper*. Cette ressemblance en a fait en quelque sorte un précurseur des *data papers* et des données qui les accompagnent, même si les aspects de contrôle scientifique des linguistes n'étaient pas aussi avancés qu'aujourd'hui.
En effet, la mise en avant récente des données scientifiques dites FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) et des *data papers* a amené la communauté scientifique à préciser la manière de partager les corpus de langage oral en fournissant des moyens (centres de dépôt et diffusion, outils), des formats (métadonnées, TEI), des méthodes et des usages, qui ont rendu plus aisée et plus fréquente la diffusion de ces corpus qui sont aujourd'hui beaucoup plus faciles à trouver qu'auparavant. Cette évolution bénéfique pour la recherche scientifique reste encore à généraliser dans le domaine des corpus de langage oral pour créer de véritables revues de *data paper* qui puissent, associées aux données FAIRisées, offrir à ce domaine de recherche les mêmes services que ceux que l'on trouve déjà dans d'autres disciplines.

Les *data papers* dans l'écosystème Recherche Data Gouv

Joachim Schöpfel et Christine Kosmopoulos

Ce chapitre décrit la mise en place de l'écosystème Recherche Data Gouv en France et analyse la place des *data papers* dans ce nouvel environnement (ateliers de la donnée, centres de références thématiques, etc.). En particulier, il évoque la génération d'une ébauche de *data paper* par l'entrepôt national et la connexion avec les revues scientifiques.

Vers un écosystème d'écriture et d'édition avec les données

Nicolas Sauret, Stéphane Pouyllau et Mélanie Bunel

En considérant une certaine convergence des « écritures scientifiques » au sens large, qui se concrétise par exemple dans les formats *data paper* et *executable paper*, un écosystème intégré et transparent dans ses échanges de données ouvre la voie à une appropriation collective des méthodologies et des bonnes pratiques numériques, que ce soit en termes d'écriture et d'édition scientifiques, ou en termes de gestion et de traitement des données de la recherche. Dans ce chapitre, nous rappelons en première partie quelques éléments de contexte et d'évolution actuelle sur les infrastructures de recherche en SHS, avant de présenter dans une seconde partie les fondements d'un écosystème de recherche en devenir à travers les interactions possibles ou existantes entre cinq composantes de l'infrastructure Huma-Num. Nous discutons enfin les principes envisagés pour un tel écosystème et leurs implications sur les pratiques de recherche et sur les communautés.

HTR-United : un écosystème pour une approche mutualisée de la transcription automatique des écritures manuscrites

Alix Chagué, Thibault Clérice et Laurent Romary

La reconnaissance des écritures manuscrites (HTR pour *Handwritten Text Recognition*) est un procédé informatique qui vise à obtenir un équivalent de texte numérique à partir de l'image d'un document physique comportant du texte manuscrit. S'appuyant sur GitHub, HTR-United invite la communauté

des utilisateurs à décloisonner les données issues de différentes plateformes HTR afin de réduire leur coût de production. Cette solution propose un modèle opérationnel susceptible d'offrir un cadre pour la construction de *data papers* pour l'HTR, voire les prémices d'une standardisation pour ce genre de publication.

De l'entrepôt de données aux *data papers*. Retour sur l'expérience de Data Sciences Po

Alina Danciu, Anna Egea, Guillaume Garcia et Cyril Heude

Ce chapitre restitue la manière dont la mise en œuvre d'un entrepôt Dataverse mutualisé – Data Sciences Po – nous a poussés à investir la question des *data papers*. Le compte rendu croise les expériences des différents acteurs (laboratoire, service transverse, centre de données) concernés. Nous remettons en contexte les dispositifs qui ont précédé la mise en place de Data Sciences Po, puis nous revenons sur l'expérience de mise en place de cet entrepôt, ainsi que sur les coûts importants nécessaires pour sa maintenance et son animation. Nous soulignons les problèmes que nous rencontrons plus particulièrement pour favoriser la documentation des données, quelles que soient leurs modalités de dépôt – auto-dépôt accompagné ou curation des données assuré pour le compte des déposants. Inciter les chercheuses et chercheurs à auto-documenter leurs jeux de données, via les *data papers*, est un enjeu central, afin de faire face aux contraintes d'accompagnement disponible (notamment le manque de moyens humains). Nous présentons enfin les solutions que nous commençons à mettre en place pour développer la pratique de rédaction de *data papers*.

Le *data paper* appliqué à la biodiversité : standards, outils et processus mis en œuvre pour démocratiser le concept dans la communauté de la bio-informatique

Sophie Pamerlon

Le GBIF (Global Biodiversity Information Facility) et Pensoft Publishers ont repris et développé le concept de *data paper* en 2011, afin de faciliter le partage et la réutilisation des données dans les domaines de la biodiversité et des sciences environnementales. Ce partenariat, via l'utilisation d'outils et de standards d'échanges de données et de métadonnées communs ou compatibles, a mis en place un processus de génération semi-automatisé de *data papers*, dans le but de mettre en valeur les données et le temps passé par leurs gestionnaires à les traiter et à les partager suivant les bonnes pratiques FAIR en vigueur dans les communautés scientifiques.

Perspectives

Christine Kosmopoulos, Victoria Le Fourner et Joachim Schöpfel

En guise de conclusion, le dernier chapitre revient sur quelques aspects essentiels : le rôle des *data papers* dans l'écosystème émergent des infrastructures des données de recherche, les différentes pratiques en matière de rédaction des *data papers*, mais aussi sur la génération automatique d'un *data paper*, l'importance des métadonnées et le lien avec les principes FAIR. En tenant compte de la « bibliodiversité » du paysage des *data papers*, nous posons la question de l'intérêt de ce nouveau type de publication pour les SHS.

La collection Humanités numériques et science ouverte (HNSO), co-dirigée par Clarisse Bardiot et Émilien Ruiz, est financée par le Fonds national pour la science ouverte et portée par la Maison Européenne des Sciences de l'Homme et de la Société (MESHS) et les Presses universitaires du Septentrion (PUS).

Elle a pour objectif de publier en *open access* des monographies et des ouvrages collectifs ainsi que les données associées. Contribuant ainsi à l'ouverture et à la diffusion des données, la collection se veut aussi un terrain d'expérimentation et de réflexion en pratique sur ce que la science ouverte fait aux SHS.
Défendant une conception pluraliste des humanités numériques, cette collection s'adresse aux spécialistes des diverses disciplines des sciences humaines et sociales qui inscrivent leurs travaux dans une démarche empirique et accordent une attention particulière à la constitution, la structuration, l'exploitation et à la visualisation de leurs données ; sans exclusive concernant les types de sources, les méthodes employées ou les tailles de corpus mobilisés.

Cet ouvrage a été financé par le Fonds national pour la science ouverte.

ISBN : 978-2-7574-4329-3
ISSN : en cours

Ouvrage composé par
Jonas Mazot & Émilie Pouderoux

Ouvrage réalisé avec
La chaîne d'édition XML-TEI Métopes
Méthodes et outils pour l'édition structurée

Dépôt légal
février 2025

2 316e volume édité par
les Presses universitaires du Septentrion
Villeneuve d'Ascq – France